Nadia Magnenat-Thalmann, Lakhmi C. Jain, and Nikhil Ichalkaranje (Eds.)

New Advances in Virtual Humans

VISIT…

Studies in Computational Intelligence, Volume 140

Further volumes of this series can be found on our homepage: springer.com

Vol. 119. Slawomir Wiak, Andrzej Krawczyk
and Ivo Dolezel (Eds.)
Intelligent Computer Techniques in Applied Electromagnetics, 2008
ISBN 978-3-540-78489-0

Vol. 120. George A. Tsihrintzis and Lakhmi C. Jain (Eds.)
Multimedia Interactive Services in Intelligent Environments, 2008
ISBN 978-3-540-78491-3

Vol. 121. Nadia Nedjah, Leandro dos Santos Coelho
and Luiza de Macedo Mourelle (Eds.)
Quantum Inspired Intelligent Systems, 2008
ISBN 978-3-540-78531-6

Vol. 122. Tomasz G. Smolinski, Mariofanna G. Milanova
and Aboul-Ella Hassanien (Eds.)
Applications of Computational Intelligence in Biology, 2008
ISBN 978-3-540-78533-0

Vol. 123. Shuichi Iwata, Yukio Ohsawa, Shusaku Tsumoto, Ning
Zhong, Yong Shi and Lorenzo Magnani (Eds.)
Communications and Discoveries from Multidisciplinary Data, 2008
ISBN 978-3-540-78732-7

Vol. 124. Ricardo Zavala Yoe
Modelling and Control of Dynamical Systems: Numerical Implementation in a Behavioral Framework, 2008
ISBN 978-3-540-78734-1

Vol. 125. Larry Bull, Bernadó-Mansilla Ester
and John Holmes (Eds.)
Learning Classifier Systems in Data Mining, 2008
ISBN 978-3-540-78978-9

Vol. 126. Oleg Okun and Giorgio Valentini (Eds.)
Supervised and Unsupervised Ensemble Methods and their Applications, 2008
ISBN 978-3-540-78980-2

Vol. 127. Régie Gras, Einoshin Suzuki, Fabrice Guillet
and Filippo Spagnolo (Eds.)
Statistical Implicative Analysis, 2008
ISBN 978-3-540-78982-6

Vol. 128. Fatos Xhafa and Ajith Abraham (Eds.)
Metaheuristics for Scheduling in Industrial and Manufacturing Applications, 2008
ISBN 978-3-540-78984-0

Vol. 129. Natalio Krasnogor, Giuseppe Nicosia, Mario Pavone
and David Pelta (Eds.)
Nature Inspired Cooperative Strategies for Optimization (NICSO 2007), 2008
ISBN 978-3-540-78986-4

Vol. 130. Richi Nayak, Nikhil Ichalkaranje
and Lakhmi C. Jain (Eds.)
Evolution of the Web in Artificial Intelligence Environments, 2008
ISBN 978-3-540-79139-3

Vol. 131. Roger Lee and Haeng-Kon Kim (Eds.)
Computer and Information Science, 2008
ISBN 978-3-540-79186-7

Vol. 132. Danil Prokhorov (Ed.)
Computational Intelligence in Automotive Applications, 2008
ISBN 978-3-540-79256-7

Vol. 133. Manuel Graña and Richard J. Duro (Eds.)
Computational Intelligence for Remote Sensing, 2008
ISBN 978-3-540-79352-6

Vol. 134. Ngoc Thanh Nguyen and Radoslaw Katarzyniak (Eds.)
New Challenges in Applied Intelligence Technologies, 2008
ISBN 978-3-540-79354-0

Vol. 135. Hsinchun Chen and Christopher C. Yang (Eds.)
Intelligence and Security Informatics, 2008
ISBN 978-3-540-69207-2

Vol. 136. Carlos Cotta, Marc Sevaux
and Kenneth Sörensen (Eds.)
Adaptive and Multilevel Metaheuristics, 2008
ISBN 978-3-540-79437-0

Vol. 137. Lakhmi C. Jain, Mika Sato-Ilic, Maria Virvou,
George A. Tsihrintzis, Valentina Emilia Balas
and Canicious Abeynayake (Eds.)
Computational Intelligence Paradigms, 2008
ISBN 978-3-540-79473-8

Vol. 138. Bruno Apolloni, Witold Pedrycz, Simone Bassis
and Dario Malchiodi
The Puzzle of Granular Computing, 2008
ISBN 978-3-540-79863-7

Vol. 139. Jan Drugowitsch
Design and Analysis of Learning Classifier Systems, 2008
ISBN 978-3-540-79865-1

Vol. 140. Nadia Magnenat-Thalmann, Lakhmi C. Jain
and Nikhil Ichalkaranje (Eds.)
New Advances in Virtual Humans, 2008
ISBN 978-3-540-79867-5

Nadia Magnenat-Thalmann
Lakhmi C. Jain
Nikhil Ichalkaranje
(Eds.)

New Advances in Virtual Humans

Artificial Intelligence Environment

Springer

Professor Dr. Nadia Magnenat-Thalmann
Vice-Rector University of Geneva
Director of MIRALab / C.U.I.
University of Geneva
24, rue General Dufour, 1211, Geneve-4
Switzerland

Dr. Nikhil Ichalkaranje
KES Centre
School of Electrical and Information Engineering
University of South Australia
Adelaide, Mawson Lakes Campus
South Australia SA 5095
Australia

Professor Dr. Lakhmi C. Jain
Knowledge-Based Engineering
Founding Director of the KES Centre
School of Electrical and Information Engineering
University of South Australia
Adelaide, Mawson Lakes Campus
South Australia SA 5095
Australia
E-mail: Lakhmi.jain@unisa.edu.au

ISBN 978-3-540-79867-5 e-ISBN 978-3-540-79868-2

DOI 10.1007/978-3-540-79868-2

Studies in Computational Intelligence ISSN 1860949X

Library of Congress Control Number: 2008926083

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1
springer.com

Preface

In the early eighties, Research on Virtual Humans (VH) began mainly in the USA and Canada. For the next 20 years, the main research in this area attempted to simulate realistic virtual humans. The intention was to simulate the appearance of a human that appears to be real. Initially an attempt was made to model the face and body. Then the next area of research was to make an adaptative skeleton for any VH size. This was developed so as to allow animation of the VH. In early nineties further research on the hair and clothes have been further addressed. At the end of last decade, it was possible to produce a real looking VH. This was successful but of limited success in real time.

Recently the main concern was to work on the behaviour and the mind of the VH. VH are expected to recognize others, to have memory, and to be able to speak and understand what is said and to have emotions. It assembles the traditional approach of artificial intelligence and the VH. The methods required should be adapted to the specificity of the VH. If we simulate gesture, speech or emotions, it should not be done by using geometric changes which are controlled by the programmer or user. It should be done at an upper level similar to a brain. In short until now, we have been working on the body and face. We are now linking them to a brain model. The brain, by using AI techniques, will supervise the actions of the body or face instead of animator. The VH should be able to react to events, to make plans, to recognize others, to behave socially and be socially adept.

In this book, various aspects of cognitive and emotional behaviour of VH is described. In chapter one, a state of the art introduction to VH is presented and the associated research is given. In Chapter 2, cognitive and emotions processes are described. A comprehensive context model for multi-party interactions with the VH is given in the next chapter. Finally, it is very important to model the socializing of groups of virtual humans. This is discussed in Chapter 4. The automatic modelling of expressions for VH is described in Chapter 5. The last chapter gives a case study of an intelligent kios avatar and its usability.

This book gives examples of some advances that enable VH to behave intelligently. It provides an overview of these research problems and some unsolved problems.

We appreciate the editorial support of Scientific Publishing Services Pvt. Ltd.

Nadia Magnenat-Thalmann
Lakhmi C. Jain
Nikhil Ichalkaranje

Contents

1

Innovations in Virtual Humans

Nadia Magnenat-Thalmann[1] and Daniel Thalmann[2]

[1] MIRALab, University of Geneva, Battelle, Carouge, Switzerland
thalmann@miralab.unige.ch
http://www.miralab.unige.ch
[2] VRlab, EPFL, Station 14, CH 1015 Lausanne, Switzerland
Daniel.Thalmann@epfl.ch
http://vrlab.epfl.ch

Abstract. This Chapter presents the main concepts of Virtual Humans. After a brief introduction, it gives a survey of the history of Virtual Humans starting from the earlier models of the sixties. Then, the modeling of human bodies is explained as well as the way body deformations can be handled. A large Section is dedicated to the animation of the skeleton with a description of main motion control methods like motion capture, parametric key-frame animation, or inverse kinematics. Motion retargeting, locomotion, and grasping techniques are also explained in details. Facial animation is presented in a specific Section as the methods are well different from the methods used for the body. A Section on autonomous characters describes properties like emotions, memory, or motivations. Finally, the last Section presents the main challenges when animating crowds.

1.1 Introduction

1.1.1 Why Virtual Humans?

Virtual Humans is a research domain concerned with the simulation of human beings on computers. It involves representation, movement and behavior. There is a wide range of applications: film and TV productions, ergonomic and usability studies in various industries (aerospace, automobile, machinery, furniture etc.), production line simulations for efficiency studies, clothing industry, games telecommunications (clones for human representation), medicine, etc. These applications have various needs. A medical application might require an exact simulation of certain internal organs; film industry requires highest esthetic standards, natural movements and facial expressions; ergonomic studies require faithful body proportions for a particular population segment and realistic locomotion with constraints; etc.

Sheer complexity of the simulated subject, i.e. the human body, combined with the multitude of applications and requirements, make Virtual Humans a vast research domain comprising numerous research topics:

- *Anatomy and geometry*, dealing with creation of human shape in 3D graphics, with methods ranging from point-to-point digitizing from clay models through various software tools for geometry deformation and modeling to laser 3D scanners.
- *Hair and skin representation and rendering*

N. Magnenat-Thalmann et al. (Eds.): New Adv. Virt. Hum.: Arti. Intel., SCI 140, pp. 1–41, 2008.
springerlink.com

- *Skeleton animation*, or animation of joint angles of the skeleton structure defining the articulated body and consisting of segments (representing limbs) and joints (representing degrees of freedom). Main methods of skeleton animation are parametric keyframe animation, direct and inverse kinematics (possibly with constraints), direct and inverse dynamics.
- *Body surface animation and deformation*, trying to simulate natural-looking movement and deformation of visible body surface with respect to the movement of the underlying skeleton structure.
- *Hand animation and deformation*
- *Facial animation*, playing an essential role for human communication. Two main stream facial animation research exist: parametrized models and muscle models.
- *Walking*, i.e. generating natural-looking walking motion based on a given trajectory and velocity.
- *Obstacle avoidance*, finding optimal trajectory for walking while avoiding obstacles
- *Grasping*, i.e. producing appropriate arm and hand motion to reach for and grab an object.
- *Behavioral animation*, striving to give more character and personality to the animation, thus making it look more natural than mechanics-based animations

An ultimate research objective is the simulation of Virtual Worlds inhabited by a Virtual Human Society, where Virtual Humans will co-operate, negotiate, make friends, communicate, group and ungroup, depending on their likes, moods, emotions, goals, fears, etc. But such interaction and corresponding groups should not be programmed. Behaviour should emerge as a result of a multi-agent system sharing a common environment, in our case, sharing a Virtual Environment. For example in a panic situation, we do not model the group behaviour, because each human reacts differently depending for example on his/her level of fear. If we model the individual entity, there will be groups of different behaviours (not programmed explicitly) as result of the interaction of common individual behaviours. Simulations consist of group of autonomous Virtual Human agents existing in dynamic 3D Virtual Environment. Virtual Humans have some natural needs like hunger, tiredness, etc. which guide selection of their behaviours. In order to behave in believable way, these agents also have to act in accordance with their surrounding environment, be able to react to its changes, to the other agents and also to the actions of real humans interacting with the virtual world. Behaviour models should be developed that are simple enough to allow for real-time execution of group of agents, yet still sufficiently complex to provide interesting behaviours. Virtual Humans have their own motivations and needs, are able to sense and explore their environment, and their action selection mechanism determine suitable actions to take at any time. For this purpose, architecture allowing merging of individuals' artificial life simulation with the current multi-agent model is developed.

1.1.2 History of Virtual Humans

Ergonomic analysis provided some of the earliest applications in computer graphics for modeling a human figure and its motion. One of the earliest figures used for ergonomic analysis was William Fetter’s Landing Signal Officer (LSO), developed for Boeing in 1959 [1]. The seven jointed "First Man", used for studying the instrument panel of a Boeing 747, enabled many pilot motions to be displayed by articulating the figure’s pelvis, neck, shoulders, and elbows. Possibly the first use of computer graphics in commercial advertising took place in 1970 when this figure was used for a Norelco television commercial. The addition of twelve extra joints to "First Man" produced "Second Man". This figure was used to generate a set of animation film sequences based on a series of photographs produced by Muybridge [2]. "Third Man and Woman" was a hierarchical figure series with each figure differing by an order of magnitude in complexity. These figures were used for general ergonomic studies. The most complex figure had 1000 points and was displayed with lines to represent the contours of the body. In 1977, Fetter produced "Fourth Man and Woman" figures based on data from biostereometric tapes. These figures could be displayed as a series of colored polygons on raster devices.

Cyberman (Cybernetic man-model) was developed by Chrysler Corporation for modeling human activity in and around a car [3]. Although he was created to study the position and motion of car drivers, there is no check to determine whether his motions are realistic and the user was responsible for determining the comfort and feasibility of a position after each operation. It is based on 15 joints; the position of the observer is predefined.

Combiman (Computerized biomechanical man-model) was specifically designed to test how easily a human can reach objects in a cockpit [4]. Motions have to be realistic and the human can be chosen at any percentile from among three-dimensional human models. The vision system is very limited. Combiman is defined using a 35 internal-link skeletal system. Although the system indicated success or failure with each reach operation, the operator was required to determine the amount of clearance (or distance remaining to the goal).

Boeman was designed in 1969 by Boeing Corporation [5]. It is based on a 50th-percentile three-dimensional human model. He can reach for objects like baskets but a mathematical description of the object and the tasks is assumed. Collisions are detected during Boeman’s tasks and visual interferences are identified. Boeman is built as a 23-joint figure with variable link lengths.

Sammie (System for Aiding Man Machine Interaction Evaluation) was designed in 1972 at the University of Nottingham for general ergonometric design and analysis [6]. This was, so far, the best parameterized human model and it presents a choice of physical types: slim, fat, muscled, etc. The vision system was very developed and complex objects can be manipulated by Sammie, based on 21 rigid links with 17 joints. The user defined the environment by either building objects from simple primitives, or by defining the vertices and edges of irregular shaped objects. The human model was based on a measurement survey of a general population group.

Buford was developed at Rockwell International in Downey, California to find reach and clearance areas around a model positioned by the operator [5]. The figure represented a 50th-percentile human model and was covered by CAD-generated polygons. The user could interactively design the environment and change the body position and limb sizes. However, repositioning the model was done by individually moving the body and limb segments. He has some difficulty in moving and has no vision system. Buford is composed of 15 independent links that must be redefined at each modification.

In 1971 Parke produces a representation of the head and face in the University of Utah, and three years later advances in good enough parametric models to produce a much more realistic face [7].

There was another popular approach based on volume primitives. Several kinds of elementary volumes have been used to create such models e.g. cylinders by Poter and Willmert [8] or ellipsoids by Herbison-Evans [9]. Designed by Badler and Smoliar [10], Bubbleman is a three-dimensional human figure consisting of a number of spheres or bubbles. The model is based on overlap of spheres, and the appearance (intensity and size) of the spheres varies depending on the distance from the observer. The spheres correspond to a second level in a hierarchy; the first level is the skeleton.

In the early 1980's, Tom Calvert, a professor of kinesiology and computer science at Simon Fraser University, attached potentiometers to a body and used the output to drive computer animated figures for choreographic studies and clinical assessment of movement abnormalities. To track knee flexion, for instance, they strapped a sort of exoskeleton to each leg, positioning a potentiometer alongside each knee so as to bend in concert with the knee. The analog output was then converted to a digital form and fed to the computer animation system. Their animation system used the motion capture apparatus together with Labanotation and kinematic specifications to fully specify character motion [11].

In the beginning of the Eighties, several companies and research groups produced short films and demos involving Virtual Humans. Information International Inc, commonly called Triple-I or III, had a core business based on high resolution CRTs which could be used for digital film scanning and digital film output capabilities which were very advanced for the time. Around 1975, Gary Demos, John Whitney Jr, and Jim Blinn persuaded Triple I management to put together a "movie group" and try to get some "Hollywood dollars". They created various demos that showed the potential for computer graphics to do amazing things, among them a 3D scan of Peter Fonda's head, and the ultimate demo, Adam Powers, or "the Juggler".

In 1982, in collaboration with Philippe Bergeron, Nadia Magnenat-Thalmann and Daniel Thalmann produced Dream Flight, a film depicting a person (in the form of an articulated stick figure) transported over the Atlantic Ocean from Paris to New York (see Figure 1.1). The film was completely programmed using the MIRA graphical language, an extension of the Pascal language based on graphical abstract data types. The film got several awards and was shown at the SIGGRAPH ‘83 Film Show.

A very important milestone happens in 1985, when the Film "Tony de Peltrie" uses for the first time facial animation techniques to tell a story. The same year, the Hard Woman video for the Mick Jagger's song was developed by Digital Productions and showed a nice animation of a woman. "Sexy Robot" was created in 1985 by Robert Abel & Associates as a TV commercial and imposes new standards for the movement of the human body (it introduces motion control).

In 1987, the Engineering Society of Canada celebrated its 100th anniversary. A major event, sponsored by Bell Canada and Northern Telecom, was planned for the Place des Arts in Montreal. For this event, Nadia Magnenat-Thalmann and Daniel Thalmann simulated Marilyn Monroe and Humphrey Bogart meeting in a cafe in the old

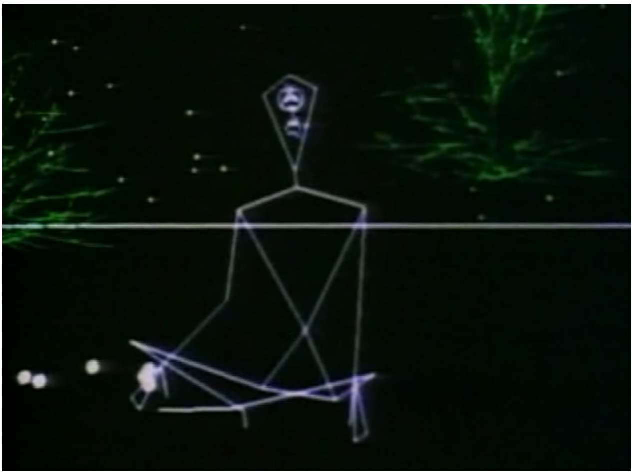


Fig. 1.1. Dream Flight

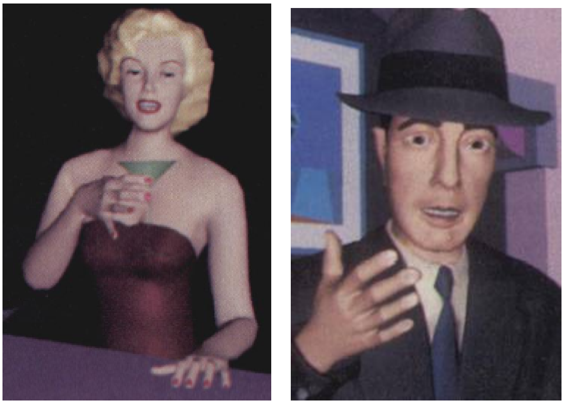


Fig. 1.2. Rendez-vous à Montréal

town section of Montreal. The development of the software and the design of the 3D characters (now capable of speaking, showing emotion, and shaking hands) became a full year's project for a team of six. Finally, in March l987, the actress and actor were given new life as virtual humans. Figure 1.2 shows both actors.

In 1988 "Tin Toy" is a winner of the first Oscar (as Best Animated Short Film) to a piece created entirely within a computer. The same year, deGraf/Wahrman developed "Mike the Talking Head" for Silicon Graphics to show off the real-time capabilities of their new 4D machines. Mike was driven by a specially built controller that allowed a single puppeteer to control many parameters of the character's face, including mouth, eyes, expression, and head position. The Silicon Graphics hardware provided real-time interpolation between facial expressions and head geometry as controlled by the performer. Mike was performed live in that year's SIGGRAPH film and video show.

The live performance clearly demonstrated that the technology was ripe for exploitation in production environments.

In 1989, Kleiser-Walczak produced Dozo, a (non-real-time) computer animation of a woman dancing in front of a microphone while singing a song for a music video. They captured the motion using an optically-based solution from Motion Analysis with multiple cameras to triangulate the images of small pieces of reflective tape placed on the body. The resulting output is the 3-D trajectory of each reflector in the space.

In 1989, in the film "The Abyss", there is a sequence where the watery pseudopod acquires a human face. This represents an important step for future synthetic characters. In 1989, Lotta Desire, actress of "The Little Death" and "Virtually Yours" establishes new accomplishments. Then, the "Terminator II" movie marked in 1991 a milestone in the animation of synthetic actors mixed with live action. In the nineties, several short movies were produced, the most well-known is “Geri's Game” from Pixar which awarded the Academy Award for Animated Short.

This is also at this time that the Jack software package was developed at the Center for Human Modeling and Simulation at the University of Pennsylvania, and was made commercially available from Transom Technologies Inc. Jack provided a 3D interactive environment for controlling articulated figures. It featured a detailed human model and included realistic behavioral controls, anthropometric scaling, task animation and evaluation systems, view analysis, automatic reach and grasp, collision detection and avoidance, and many other useful tools for a wide range of applications."

In the 1990s, the emphasis has shifted to real-time animation and interaction in virtual worlds. Virtual Humans have begun to inhabit virtual worlds and so have we. To prepare our place in the virtual world, we first develop techniques for the automatic representation of a human face capable of being animated in real time using both video and audio input. The objective is for one's representative to look, talk, and behave like oneself in the virtual world. Furthermore, the virtual inhabitants of this world should be able to see our avatars and to react to what we say and to the emotions we convey.

The virtual Marilyn and Bogart are now 17 years old. The Virtual Marilyn has acquired a degree of independent intelligence; she even played in 1997 the autonomous role of a referee announcing the score of a real-time simulated tennis match on a virtual

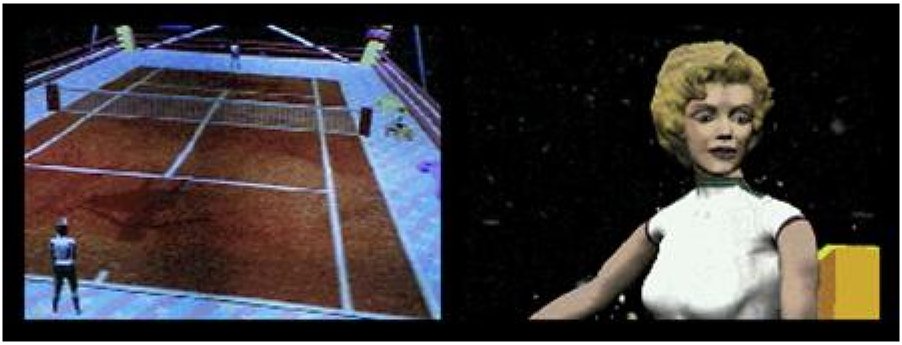


Fig. 1.3. Marilyn as a referee for tennis

court, contested by the 3D clones of two real players situated as far apart as Los Angeles and Switzerland [12]. Figure 1.3 shows the tennis play and Marilyn as a referee.

1.2 Human Body Modeling

In 1987, Magnenat-Thalmann et al. [13] developed joint-dependent local deformation (JLD) operators to deform human hand model's surface for animation. Each JLD operator is affecting its uniquely defined domain and its value is determined as a function of angular values of the joints under the operator. Skeletal structure of the hand is used to implement the joints of the model. Later Magnenat-Thalmann et al. [16] used the same method on full body animation with the full skeletal joints. An example is shown in Figure 1.2.

Based on Lasseter [14]'s layered animation approach, Chadwick [15] developed layered model construction method to achieve the visual realism. Aim was to design complex body models by parametric constraints that are affecting the layered structure of the model. Critter, which is a prototype of the layered animated character modeling system, is designed to achieve this aim. Critter system supports four layers for modeling:

1) Behaviour layer that specifies the motion of the model.
2) Skeleton layer managing the articulation.
3) Muscle and fatty tissue layer.
4) Skin layer.

Skeleton layer is the base for other layers. It provides the articulation hierarchy from which the other layers built for final deformation. According to the Critter system, skeleton data includes:

1) Hierarchy of robotic manipulators, which are hierarchy of joints and links.
2) Denevit&Hartenberg(D&H) joint parameters.
3) Joint angle constraints and physical attributes.

In robotics literature, D&H parameters [16] define the hierarchy of the joints, links and their behaviour.

Muscle and fatty tissue layer is a mapping from skin data to the underlying skeleton layer. Free form deformation method is used for modeling this layer. Set of prototype deformation operators are provided for skin deformation through muscle abstraction. Each muscle is represented with a pair of FFD cube with 7 planes of control points where planes are orthogonal to the link axis.

Overall deformations are based on dynamic and kinematic constraints. Kinematic constraints are skeletal state of the underlying joints that are providing squash and stretch behaviours. The kinesiology literature [17] is using elasticity and contractility properties to define muscle and joint action characteristics. According to these properties, Chadwick developed a set of algorithms to model the flexor and extensor, tendon deformation for each frame. After skeletal motion computation, dynamic deformation is applied on the final skin mesh by using mass-spring approach on the pre-computed FFD cube.

One of the earlier researches about anthropometric modeling of the human body model is introduced by Azuola [18]. First the human body model is segmented into groups according to the synovial joints. Deformation on the corresponding joint is constrained with its degree of freedom (DOF). According to the anthropometric measurement database, anthropometrically segmented body model deformed with FFD methods. Azuola developed a system called JAKE that performs shape estimation segmentation and joint determination. Output of the system is used to generate virtual body constructed with deformable segments. Each segment is a geometric primitive connected by joints which have less then four degree of freedom. Polyhedral human body model segments are uniformly and non-uniformly scaled to construct different size bodies based on SASS system which is an anthropometric body measurement spreadsheet like system.

Shen [19] used a new approach for body model representation. Instead of polygonal representation, model divided into slices. Each slice is defined with parametric curves namely b-splines. Depending on the neighbouring joints distance and the normal vectors of the slices, collision prevented. Radius of each contour is scaled to achieve the muscular deformation effect.

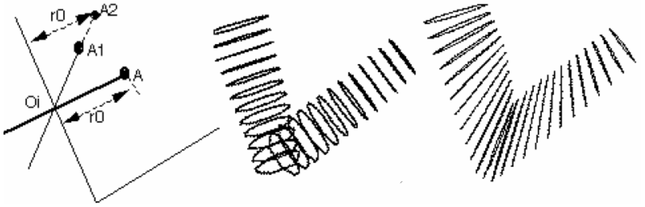


Fig. 1.4. Contour representation of arm and its deformed state, image by Shen [19]

Dividing body model into slices is proceeded by first segmenting it into six groups namely; Torso, hip, left and right arms and legs separately. Using ray casting method, from center of the segment to the skin rays are projected. Resulting skin ray intersection points are then used as control points of b-spline curve to represent each slice. According to the angle between to link, slices on the joint region are rotated with respect to their center. Figure 1.4 represents the deformation process on the join part of the model.

Later on Shen extended the contour based representation of body model with metaballs [20]. Like cylindrical representation, metaballs are used for smooth and detailed modeling. Instead of working in 3D space, model is divided into slices to reduce the complexity into 2D space. By this way, surface construction problem is reduced to curve analyses. Starting with a few metaballs, details of each slice is achieved by adding, changing the metaball parameters.

Wilhelms [21] used anatomical modeling techniques to produce realistic muscle deformation effects on the model surface. Multi layered model is defined by skeleton, muscle and the skin. Underlying layers are represented with ellipsoids because of its simplicity and computational efficiency. Body model defined as a three structure to have a hierarchical layout. User has the flexibility to add remove new nodes (skeletons etc.) into this hierarchy. Each structure in this hierarchy has its own coordinate frame that makes the computations simpler. Each muscle between two neighboring skeleton is modeled with three ellipsoid to construct a real muscle model with two tendons and one muscle belly.

After muscle hierarchy constructed, initial skin is generated. Using iterative Newton-Raphson approach, for each skin vertex, the nearest point on the ellipse is found. Using this information, skin is anchored to the muscle. According to the motion of the muscles, skin surface is updated to produce realistic muscle effects on the body.

With much more improvement over his previous work, Wilhelms [22] developed a new method for modeling and animating the animals. Using ellipsoids or triangular mesh, underlying layers of the skin are modeled where the shape of muscles are automatically changed according to the movement of the corresponding joint. His modeling approach consists of the following steps:

1) Define the model hierarchy with its rest posture.
2) Design the underlying layers of the skin, especially muscle and skeleton layers.
3) Generate triangle mesh skin.
4) Skin vertices to underlying nearest component mapping.

After modeling a body, she uses the following steps for animation:

1) Defining the joint's motion flexibility.
2) Transform and deform the corresponding components of joints.
3) Determine new positions of the skin vertices.
4) Apply skin relaxation algorithm to achieve the force equilibrium on the skin surface.

Wilhelms modeled the skeleton and generalized tissues by using meshes and ellipsoids. This approach makes it possible to adapt the parameters of these components according to other individuals. On the other hand muscle layer is modeled with different approach. Normally muscle contraction causes the movement of underlying joints; in his approach muscle contraction is modeled according to the joint movement. In his previous work [21], ellipsoids are used for modeling the muscles. His new method shows that it is insufficiently general for modeling all types of muscles except the front arms. New approach based on deformed cylinder muscle method which provides better compromise between speed and visual realism. Deformed cylinder method is implemented by defining two origin and two insertion locations on a specific bone. These locations are parameterized according to the bounding box of the corresponding bone. This parameterization makes it possible to use the same implementation on other individuals. Each deformed cylinder divided by 8 elliptic slices to generate muscle segments. These slices are discretized into equal spaces radial points to generate a polygon. Later neighboring points on each slice are connected to generate the muscle mesh.

Once the insertions and origins of default muscle are determined by the user, deformed cylinder automatically constructed and would be altered by the user. Non-default muscles such as the ones on the upper arm that are connected to the torso are modeled with a bit different manner where more user interaction is required. During the animation, approximate muscle volume preservation is considered to deform the muscle. Even biological volume preservation is not satisfied, current and relaxed volume ratios and muscle lengths are used to recalculate the deformation. Novel contribution of his method can be summarized as skin deformation with response to the underlying layers. Since the skin is triangle mesh surface, it is displaced according to the underlying deformation. Skin's triangle mesh surface is generated by voxelization method. Discrete grid points of the body are filtered by Gaussian kernel with density function to determine if they are inside or outside the body. According to the user defined threshold, eliminated grid points are used to construct the triangular mesh. Next step is to anchor the skin vertex to the nearest underlying component. This is achieved by associating each skin vertex to a previously created nearest muscle segment. Finally series of relaxation operators based on the area of each skin surface polygon is used to achieve the simulation of elastic membrane.

Another research about musculature modeling is presented by Scheepers et al. [23]. In contrast to joint dependent deformation of a model skin, Scheepers offered anatomical deformation approach for plausible visual effects where their research is focused on skeletal muscle modeling. Structurally skeletal muscles consists of three parts namely belly, origin the stationary end and insertion the movable part. For belly part of the skeletal muscle, they used the ellipsoidal modeling. Two types of skeletal muscle model considered, first one is the widely used simple form called fusiform that behaves like straight lines and the second one with more complexity is multi-belly muscles that are modeled as tubular shaped bi-cubic patch meshes capped with elliptic hemispheres. Based on these modeling primitives, Scheepers developed a procedural language to describe an anatomical model.

While most of the anthropometric modeling techniques uses template model to generate different size models, Knopf implemented a generic spherical shape which is deformed to have the final shape of the body model specified by measurements. Idea is to define a spherical Bezier surface that is centered inside the model and then modify the control point weights of the Bezier surface to generate the final shape. Modification of the control point weights are handled by using Bernstein Basis function neural network. Operation principle of the neural network allows adaptive error minimization of the distance between the vertices of the Bezier surface and the desired one. Once the minimal distance is found final weights of the network is used as a control point weights of the generic Bezier surface that will represent the desired model surface.

Cloning human body models from photos is a different approach that is represented in body modeling field. Lee at al. [24] represent the earlier researches that utilizes photos of a person to construct 3D model. Taken three photos of a person from front, back and side view, 3D body model constructed. For face and hands, this method requires just two images. For 3D model construction, template body model with seamless texture mapping is used. Template model which is designed according to Mpeg compatible H-Anim standards is ready for animation because of initially attached skeleton. Photos of a person is used to extract the silhouette of the model and handled fully automatically. According to the boundary of the silhouette, template

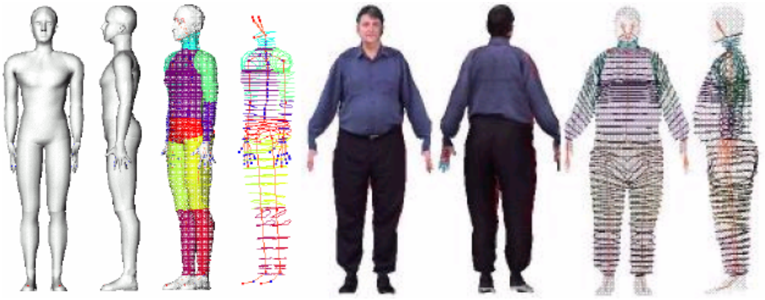


Fig. 1.5. Body model construction from orthographic images, Lee et al. [24]

model is modified to match the size of body model in photos. Boundaries of the textures that are neighboring are blended to eliminate the color difference between them. Figure 1.5 shows examples.

Modeling human bodies from a single template model has side effects during the animation. Template body model represents the model surface in a specific posture but while the model is animated underlying layers behaves different in different posture. This important aspect of human movement is modeled by Allen at al. [25]. Since the muscles, bones and other layers under the skin change their shape according to the body posture, using single template model or body scan is not enough to be used for representing an animated model. Using more then one scan or template model with different postures for each animation frame is also impossible. Allen at al. scan a body model in different key postures. For animating this model, they use scattered data interpolation method to generate smooth pass between these key frames. Resulting animation is a visually realistic model that reflects the underlying anatomic layer changes during the animation.

Since the human body scanners became widely used, it is possible to generate visually realistic models. Recent body scanner systems have the capability of capturing high resolution data along with the texture information. Apart from such benefits, these systems generate static models which require post-processing stages to let them available for further deformation and animation. In the last decade we observe key attempts to solve such problems and benefit from scanner systems high resolution data generation capabilities. An early research on this field carried out by Seo and Magnenat-Thalmann [26]. The method consists of 3 main steps for parametrically synthesize visually realistic human body models: Pre-processing the 3D scan data, function approximation of the parameterized modeler and the runtime evaluator. Together with set of scanned human body data, template model with appropriate skeleton attachment designed. For parametric deformation, specific landmarks over the template model determined. Same landmarks also specified on the scanned data set. Regarding the user specified model parameters appropriate scanned data found from the database. Finally template body model mapped on the resulting scanned data with skeleton adjustment and displacement mapping. Resulting mesh processed with refinement operator to handle irregular deformation of the template one while mapping stage.

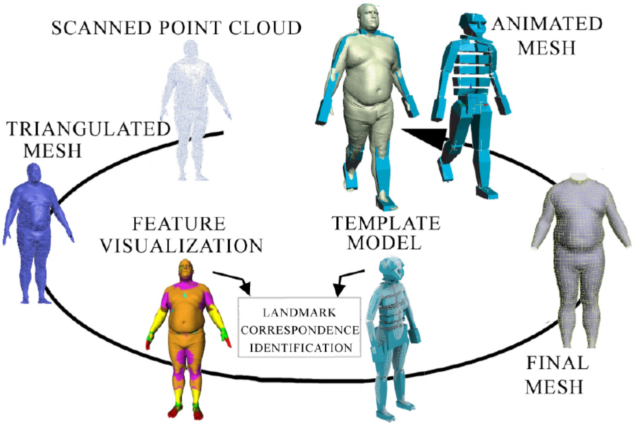


Fig. 1.6. Template based body modeling [26]

1.3 Animating the Body

1.3.1 Articulated Bodies and Virtual Characters

Virtual Humans are articulated figures modelled with multiple layers: a virtual skin is usually attached to an underlying skeleton which animates the whole body. The skeleton is a hierarchically organized set of joints, and this set depends on the animation requirements: e.g. it might be relevant for a medical surgery tool to model each vertebra of the spine, while a simpler decomposition into four joints could be sufficient for a video-game character. Of course, it is better to create morphologically-correct skeletons, but this can turn out to be quite costly. Real humans have so many degrees of freedom that virtual characters frequently omit some of them.

A skeleton can be described as a hierarchically organized set of joints, with each joint having one or more rotational degrees of freedom (DOF). It also often stores a 4D transformation matrix for representing the length and orientations of the bones, together with joints limits in order to avoid unnatural rotations. Since matrices are always expressed locally, the resulting global posture is the compositions of all previous transformation matrices in the hierarchy, starting from the root. For example, the shoulder is a joint with three DOFs and the elbow joint is its successor in the hierarchy (see Figure 1.7). If the matrix of the shoulder is modified, then the elbow moves, but its local matrix stays the same.

In an effort to standardize on a common representation for Virtual Humans the Web3D Consortium defined the Humanoid Animation Specification (H-Anim) for VRML and MPEG4. They specify a fixed topology and a naming convention for the articulated structure as well as tools for adding an outer mesh, e.g. the skin or directly cloth.

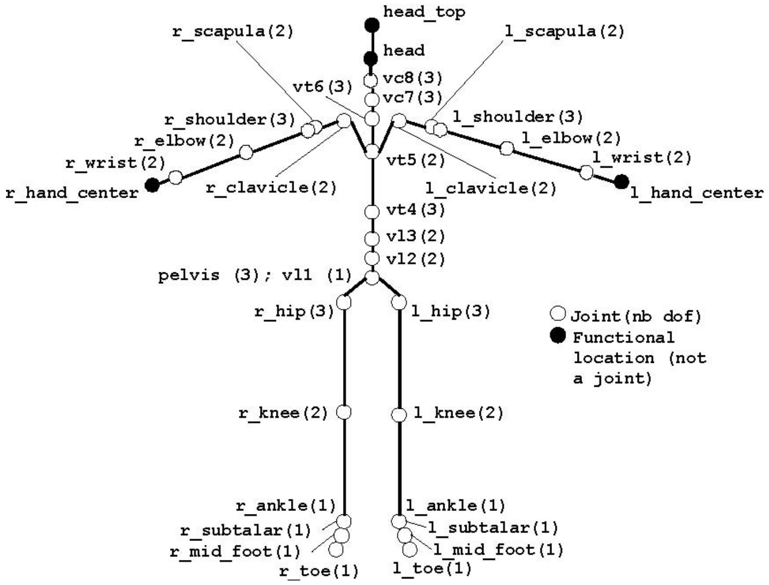


Fig. 1.7. H-ANIM skeleton

As shown in Figure 1.7, the skeleton consists of a number of segments (such as the forearm, hand and foot), which are connected to each other by the joints (such as the elbow, wrist and ankle). In order for an application to animate a Virtual Character, it needs to obtain access to the joints and alter the joint angles. The application may also need to retrieve information about such things as joint limits and segment masses. Animation of the skeleton is performed by modifying 4D matrices associated to the bones, either by synthesizing the motion, or with the help of Inverse Kinematics and Keyframing, as we are going to present in the next Sections.

1.3.2 Motion Control Methods

We will start from the classification introduced by Magnenat Thalmann and Thalmann [27] based on the method of controlling motion. A motion control method specifies how an object or a articulated bodies is animated and may be characterized according to the type of information to which it is privileged in animating the object or the character. For example, in a keyframe system for an articulated body, the privileged information to be manipulated is joint angles. In a forward dynamics-based system, the privileged information is a set of forces and torques; of course, in solving the dynamic equations, joint angles are also obtained in this system, but we consider these as derived information. In fact, any motion control method will eventually have to deal with geometric information (typically joint angles), but only geometric motion control methods are explicitly privileged to this information at the level of animation control.

The nature of privileged information for the motion control of characters falls into three categories: geometric, physical and behavioral, giving rise to three corresponding categories of motion control method.

- The first approach corresponds to methods heavily relied upon by the animator: **motion capture, shape transformation, parametric keyframe animation**. *Animated objects are locally controlled.* Methods are normally driven by geometric data. Typically the animator provides a lot of geometric data corresponding to a local definition of the motion.
- The second way guarantees a realistic motion by using physical laws, especially **dynamic simulation**. The problem with this type of animation is controlling the motion produced by simulating the physical laws which govern motion in the real world. The animator should provide physical data corresponding to the complete definition of a motion. The motion is obtained by the dynamic equations of motion relating the forces, torques, constraints and the mass distribution of objects. As trajectories and velocities are obtained by solving the equations, we may consider *actor motions as globally controlled.* Functional methods based on biomechanics are also part of this class.
- The third type of animation is called **behavioral animation** and takes into account the relationship between each object and the other objects. Moreover the control of animation may be performed at a task level, but we may also consider *the animated objects as autonomous creatures.* In fact, we will consider as a behavioral motion control method any method which drives the behavior of objects by providing high-level directives indicating a specific behavior without any other stimulus.

1.3.3 Motion Capture and Performance Animation

Performance animation or motion capture consists of measurement and recording of direct actions of a real person or animal for immediate or delayed analysis and playback. The technique is especially used today in production environments for 3D character animation. It involves mapping of measurements onto the motion of the digital character. This mapping can be direct: e.g. human arm motion controlling a character's arm motion or indirect: e.g. mouse movement controlling a character's eye and head direction. Real-time motion capture is very important in Virtual Reality, as it can provides the computer with information on the motion of the user: position and orientation of the limbs, postures, and gestures.

We may distinguish mainly two kinds of systems, optical and magnetic.

Optical Motion Capture Systems

Passive Optical Systems use markers coated with a retroreflective material to reflect light back that is generated near the cameras lens. The cameras sensitivity can be adjusted taking advantage of most cameras narrow range of sensitivity to light so only the bright markers will be sampled ignoring skin and fabric. The centroid of the marker is estimated as a position within the 2 dimensional image that is captured. The grayscale value of each pixel can be used to provide sub-pixel accuracy. Markers are attached to an actor's body and on several cameras focused on performance space. By tracking positions of markers, one can get locations for corresponding key points in the animated model, e.g. we attach markers at joints of a person and record the position of markers from several different directions. **Active optical systems** triangulate positions by illuminating one LED at a time very quickly or multiple LEDs but sophisticated software to identify them by their relative positions, somewhat akin to celestial navigation. Rather than reflecting light back that is generated externally, the markers themselves are powered to emit their own light. The use of conventional video cameras is still a dream, as it is very difficult to detect automatically the joints in 3D and to make a correct correspondence between 3D points on images captured by several cameras. This is an active area of research in Computer Vision.

Magnetic Position/Orientation Trackers and Systems

Magnetic motion capture systems require the real actor to wear a set of magnetic sensors (see Figure 1.8), as defined above, which are capable of measuring their spatial relationship to a centrally located magnetic transmitter. The position and orientation of each sensor is then used to drive an animated character. One problem is the need for synchronizing receivers. The data stream from the receivers to a host computer consists of 3D positions and orientations for each receiver. Since the sensor output has 6 degree of freedom, useful results can be obtained with two-thirds the number of markers required in optical systems; one on upper arm and one on lower arm for elbow position and angle. For complete human body motion, eleven sensors are generally needed: one on the head, one on each upper arm, one on each hand, one in the center of chest, one on the lower back, one on each ankle, and one on each foot. To calculate the rest of the necessary information, the most common way is the use of inverse kinematics. The wiring from the sensors tends to preclude extreme performance movements, but, as already seen, now there are wireless sensors on the market. The capture volumes for magnetic systems are dramatically smaller than they are for optical systems.

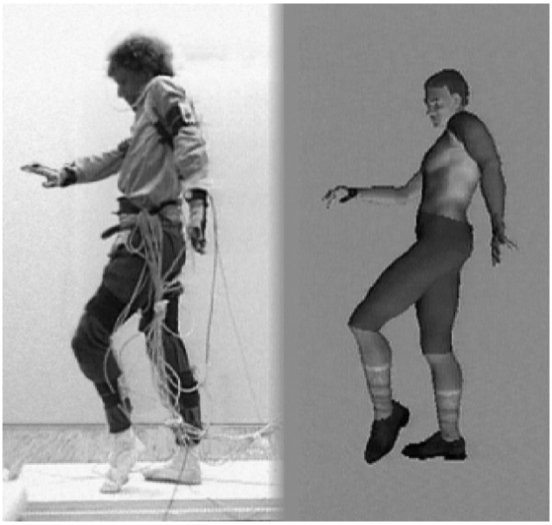


Fig. 1.8. Motion capture using magnetic sensors

1.3.4 Parametric Key-Frame Animation

This method [28] consists of the automatic generation of intermediate frames, called **in-betweens**, based on a set of key values of parameters supplied by the animator. In-between frames are obtained by interpolating the key values of the parameters and reconstructing the objects from these interpolated values. The parameters are normally spatial parameters, physical parameters and visualization parameters that decide the models' behavior.

Parametric keyframe animation is considered as a **direct kinematics method** in motion control when the interpolated parameters are defined in Joint Space of the articulated figure. Efficient and numerically well behaving methods exist for the transformation of position and velocity from Joint Space to Cartesian Space.

In both keyframe methods, linear interpolation produces undesirable effects such as lack of smoothness in motion, discontinuities in the speed of motion and distorsions in rotations. For these reasons, spline interpolation methods are used. Splines can be described mathematically as piecewise approximations of cubic polynomial functions. Two kinds of splines are very popular: interpolating splines with C1 continuity at knots, and approximating splines with C2 continuity at knots. For animation, the most interesting splines are the interpolating splines: cardinal splines, Catmull-Rom splines, and Kochanek-Bartels [29] splines.

1.3.5 Inverse Kinematics

Formally, the direct kinematics problem consists in finding the position of end point positions (e.g. hand, foot) with respect to a fixed-reference coordinate system as a function of time without regard to the forces or the moments that cause the motion. Efficient and numerically well-behaved methods exist for the transformation of position and velocity from joint-space (joint angles) to Cartesian coordinates (end of the limb).

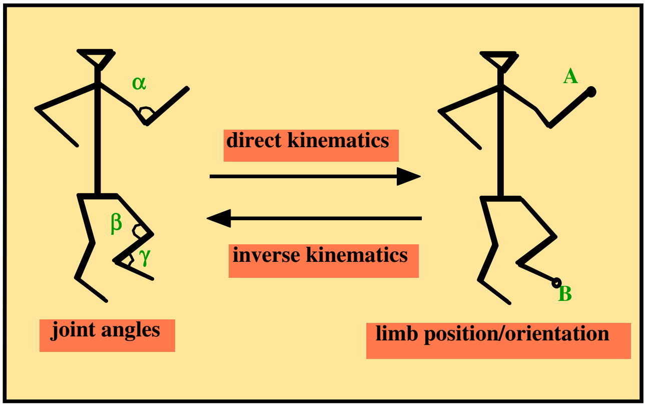


Fig. 1.9. Direct and inverse kinematics

The inverse kinematics problem is the opposite of the direct kinematics problem (see Figure 1.9). This is the determination of the joint variables given the position and the orientation of the end of the manipulator, or end effector, with respect to the reference coordinate system. It can be solved by various methods, such as inverse transform, screw algebra, dual matrices, dual quaternian, iterative, and geometric approaches. More details of each method can be found in [30].

Inversion is possible if the dimensions of Joint Space and Cartesian Space are the same. However, a general articulated structure may contain more DOF in Joint Space which are highly redundant in accomplishing tasks. The inversion is not always possible. The solution is the first order approximation of the system: to linearize the direct geometric model. As a consequence of the linearization, the solution's validity of inverse kinematics is limited to the neighborhood of the current state and, as such, any desired motion has to comply with the hypothesis of small movements.

1.3.6 Motion Retargeting

As we have seen in Section 1.3.2, there is a great interest to record motion using motion capture systems (magnetic or optical), then to try to alterate such a motion to create this individuality. This process is tedious and there is no reliable method at this stage. Even if it is fairly easy to correct one posture by modifying its angular parameters (with an Inverse Kinematics engine, for instance), it becomes a difficult task to perform this over the whole motion sequence while ensuring that some spatial constraints are respected over a certain time range, and that no discontinuities arise. When one tries to adapt a captured motion to a different character, the constraints are usually violated, leading to problems such as the feet going into the ground or a hand unable to reach an object that the character should grab. The problem of adaptation and adjustment is usually referred to as the Motion Retargeting Problem. Witkin and Popovic[31] proposed a technique for editing motions, by modifying the motion curves through warping functions and produced some of the first interesting results. In a more recent paper [32], they have extended their method to handle physical elements, such as mass and gravity, and also described how to use characters with different numbers of degrees of freedom. Their algorithm is based on the reduction of the character to an abstract character which is much simpler and only contains the degrees of freedom that are useful for a particular animation. The edition and modification are then computed on this simplified character and mapped again onto the end user skeleton. Bruderlin and Williams [33] have described some basic facilities to change the animation, by modifying the motion parameter curves. The user can define a particular posture at time t, and the system is then responsible for smoothly blending the motion around t. They also introduced the notion of motion displacement map, which is an offset added to each motion curve. The Motion Retargeting Problem term was brought up by Michael Gleicher [34]. He designed a space-time constraints solver, into which every constraint is added, leading to a big optimisation problem. He mainly focused on optimising his solver, to avoid enormous computation time, and achieved very good results. Given a captured motion associated to its Performer Skeleton, Monzani et al. [35] decompose the problem of retargeting the motion to the End User Skeleton into two steps:

1. Computing the Intermediate Skeleton matrices by orienting the Intermediate Skeleton bones to reflect the Performer Skeleton posture (Motion Converter).
2. Setting the End User Skeleton matrices to the local values of the corresponding Intermediate Skeleton matrices.

The first task is to convert the motion from one hierarchy to a completely different one. An Intermediate Skeleton model is introduced to solve this, implying three more subtasks: manually set at the beginning the correspondences between the two hierarchies, create the Intermediate Skeleton and convert the movement. We are then able to correct the resulting motion and make it enforce Cartesian constraints by using Inverse Kinematics. When considering motion conversion between different skeletons, one quickly notices that it is very difficult to directly map the Performer Skeleton values onto the End User Skeleton, due to their different proportions, hierarchies and axis systems. This raised the idea of having an Intermediate Skeleton: depending on the Performer Skeleton posture, we reorient its bones to match the same directions. We have then an easy mapping of the Intermediate Skeleton values onto the End User Skeleton (Figure 1.10).

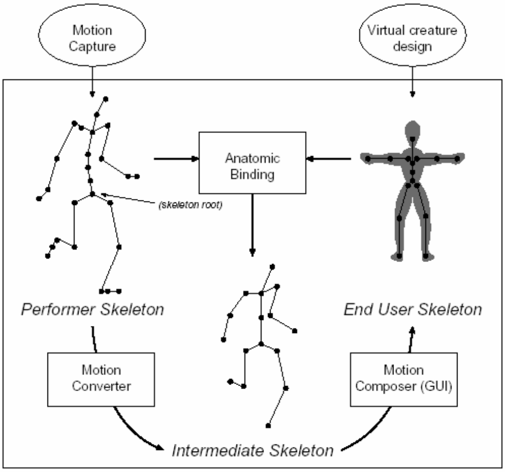


Fig. 1.10. Use of an intermediate skeleton for motion retargeting

Bindiganavale and Badler [36] also addressed the motion retargeting problem, introducing new elements: using the zero-crossing of the second derivative to detect significant changes in the motion, visual attention tracking (and the way to handle the gaze direction) and applying Inverse Kinematics to enforce constraints, by defining six sub-chains (the two arms and legs, the spine and the neck). Finally, Lee and Shin [37] used in their system a coarse-to-fine hierarchy of B-splines to interpolate the solutions computed by their Inverse Kinematics solver. They also reduced the complexity of the IK problem by analytically handling the degrees of freedom for the four human limbs. Lim and Thalmann [38] have addressed an issue of solving customers' problems when applying evolutionary computation. Rather than the seemingly more impressive approach of wow-it-all-evolved- from-nothing, tinkering with existing models can be a more pragmatic approach in doing so. Using interactive evolution, they experimentally validate this point on setting parameters of a human walk model for computer animation while previous applications are mostly about evolving motion controllers of far simpler creatures from scratch.

1.3.7 Physics-Based Animation

A great deal of work exists on the dynamics of articulated bodies [39], and efficient direct dynamics algorithms have been developed in Robotics for structures with many degrees of freedom [40]. In Computer Animation, these algorithms have been applied to the dynamic simulation of the human body [41]. Given a set of external forces (like gravity or wind) and internal forces (due to muscles) or joint torques, these algorithms compute the motion of the articulated body according to the laws of rigid body dynamics. Impressive animations of passive structures like falling bodies on stairs can be generated in this way with little input from the animator. Figure 1.11 shows an example of animation generated using dynamics. In fact, an important issue that arises in this direct or forward dynamics is how to control the model. Mathematically, forward dynamics translates into differential equations, which are typically posed as an initial-value problem; the user has little control other than setting up the initial configuration. This is exactly opposite to keyframes where the animator has the full control. Control of physics-based models remains an open research issue.

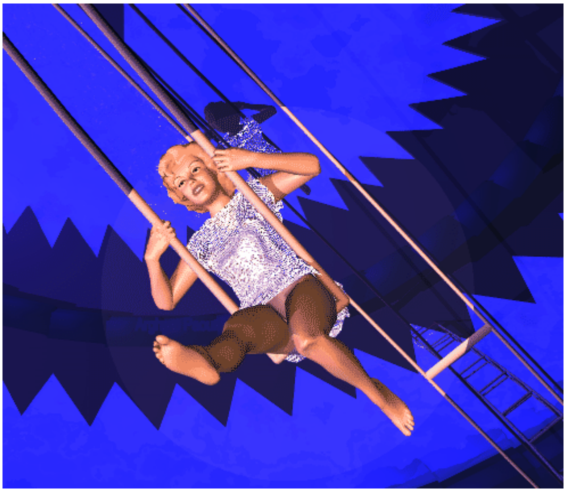


Fig. 1.11. Dynamics-based Motion

1.3.8 Locomotion

Walking has global and specific characteristics. From a global point of view, every human-walking has comparable joint angle variations. However, at a close-up, we notice that individual walk characteristics are overlaid to the global walking gait. We will take as example the walking engine described in [42].

Walking is defined as a motion where the center of gravity alternatively balances from the right to the left side. It has the following characteristics

- at any time, at least one foot is in contact with the floor, the 'single support' duration (ds).
- there exists a short instant during the walk cycle, where both feet are in contact with the floor, the 'double support' duration (dds).
- it is a periodic motion which has to be normalized in order to adapt to different anatomies.

The joint angle variations are synthesized by a set of periodic motions which we briefly mention here:

- sinus functions with varying amplitudes and frequencies for the humanoid's global translations (vertical, lateral and frontal) and the humanoid's pelvic motions (forward/backward, left/right and torsion)
- periodic functions based on control points and interpolating hermite splines. They are applied to the hip flexion, knee flexion, ankle flexion, chest torsion, shoulder flexion and elbow flexion.

More generally, many works have been dedicated to the locomotion of virtual humans [43, 44].

The keyframing technique allows an animator to specify key postures at specific key times. Using appropriate software [45], skilled designers can control the motion in detail. However, this technique is quite labor-intensive, as any motion parameter change entails the animators to modify every keyframe. Kinematics approaches generate motions from parameters such as position feet or speed value [46, 47]. Motions are generated by giving a pre-defined set of foot positions (footprints) and timing information. This data is generally computed by a motion planning technique which has to be as interactive as possible to be comfortable for animators.

Dynamics approaches aim to describe a motion by applying physics laws. For example, it is possible to use control algorithms based on finite state-machine to describe a particular motion and proportional derivative servos to compute the forces [48]. However, even if these methods produce physically correct animations, the configuration of their algorithms remains difficult. It is not easy to determine the influence of each parameter on the resulting motions. Many methods based on empirical data and bio-mechanical observations are able to generate walking [49] or running patterns [50], reactive to given user parameters. Other similar approaches take into account the environment, to walk on uneven or sloped terrains [51] or to climb stairs [52]. Despite their real-time capability, all these methods lack in realism, as the legs' motion is considered symmetrical for example.

Another class of animation techniques re-uses original motion capture data. Treated as a time-varying signal, a new motion can be generated by modifying its frequency bands [53] or its Fourier coefficients [54]. Other methods [55, 56] define each motion by B-Spline coefficients. New motions are then computed by setting weights on the various original motions and performing interpolations using polynomial and RBF (Radial Basis Function) functions. The Kovar and Gleicher [57] method wraps input motions into a data structure which ensures consistent time-warping, root alignment and constraint matching.

1.3.9 PCA-Based Method

A method from statistics, PCA [58], has recently become attractive in animation synthesis. It is used either to compress the data [59] or to emphasize similarities between input data [60] in order to generate motion according to control parameters such as age or gender. In this section, we introduce a model allowing efficient generation of a whole locomotion sequence only at each high-level parameters update. To illustrate our methodology, we use two examples of locomotion: walking and running.

The first step consists in creating a motion database; for example, we use an optical motion capture system and a treadmill to record five subjects differing in age and gender (two women and three men). The physical parameter speed of the various sequences varies from 3.0 km/h to 7.0 km/h, by increments of 0.5 km/h, in the case of walking, and from 6.0 km/h to 12.0 km/h, by increments of 1.0 km/h, for running. The sequences were then segmented into cycles (one cycle includes two steps, starting at right heel strike), and four of them have been selected. These cycles are aligned to an identical locomotion direction, converted to joint angle space (represented by axis-angles, according to the standard H-ANIM skeleton [61]), and finally normalized, so that each sequence is represented by the same number of samples. In addition, a standing (neutral) position sequence of each subject has been inserted to represent the speed value 0 km/h. Consequently, the database is composed of 180 walking cycles and 140 running cycles.

In practice, a person's posture, or body pose, in a given keyframe can be defined by the position and orientation of a root node and a vector of joint angles. A motion can then be represented by an angular motion vector μ, which is a set of such joint angle vectors measured at regularly sampled intervals.

As computing the entire locomotion sequence is time consuming, PCA technique [58] is applied, drastically reducing the dimension of the input motion capture data space. The resulting space, referred to as the main PCA, is computed with the input motion matrix M composed of all motion vectors μ from our database with k subjects. To center this space with respect to the whole data set, we define μ_0 as an average vector of all n motion vectors. The basis vectors describing this space are the m first orthogonal PC's (Principal Components) necessary to compute an approximation of the original data. Let $\alpha = (\alpha1; \alpha2; : : : ; \alpha m)$ be a coefficient vector and $E = (e_1; e_2; : : : ; e_m)$ a vector matrix of the first PC's (or eigenvectors) of M, a motion θ can be expressed as:

$$\theta \cong \theta_0 + \sum_{i=1}^{m} \alpha_i e_i = \theta_0 E \qquad (3.1)$$

As mentioned, the purpose of this PCA is to reduce the dimensionality of the input data, a very important aspect for real-time purposes. To generate a new and entire motion, blending technique could be applied on various α, according to three high-level parameters: personification vector p, where pi is the weight for the i subject, type of locomotion T (walk or run) and speed S. In theory, a linear interpolation between coefficient vectors is performed, while original data are non-linear (axis-angle). This ambiguity can be solved by applying the method presented in [62]. However, in practice, the method here is intended for locomotion, where the different motions show small variations between postures, and mostly in the sagittal plane, therefore allowing linear interpolation.

Unfortunately, blending is not appropriate for motion extrapolation. As the goal of a locomotion engine is to propose extrapolation of physical parameters, i.e speed in our example, an alterative technique has to be applied. A complete method, explained proposes a hierarchical structure of PCA spaces that first help to classify the motions, and second allow a linear least square in a very low dimension, instead of in the main PCA space.

Data are then retargeted to different human sizes from those captured, and there is a process, based on motion analysis, to unwarp the normalized data. To produce animation adaptable to any kinds of virtual humans, the generalization of the heterogeneous input data we used is an important aspect. Indeed, we captured various subjects, not only with differences in the style of motion, but also, and more importantly, differences in size. Figure 1.12 shows an example.

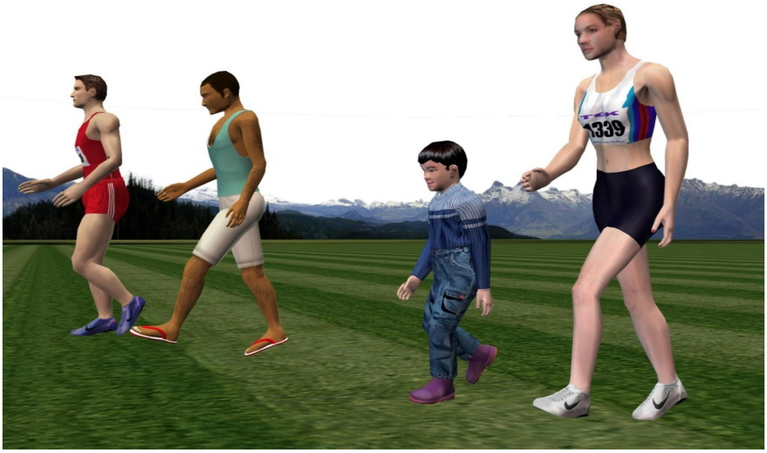


Fig. 1.12. Walking generated using PCA

1.3.10 Grasping

Grasping is perhaps the most important and complicated motion that manipulation of objects involves. The difficulty comes not only from properly “wrapping” the fingers around the object, but also from the fact that the grasp must be suitable for the intended manipulation. In [63], a classification of the hand postures commonly used for grasping in manufacturing tasks is given. One of the earlier works on grasping that uses this classification to automatically generate grasping motions is the approach described in [64, 65]. This approach is based on three steps:

- **Heuristic grasping decision** based on a grasp taxonomy [65] (see Figure 1.13a)
- **Inverse kinematics to find the final arm posture**

- **Use of virtual sensors** A sensor is activated for any collision with other objects or sensors. Here we select sphere sensors for their efficiency in collision detection. Figure 1.13b-c shows the multi-sensors and Figure 1.13d a cube grasped.

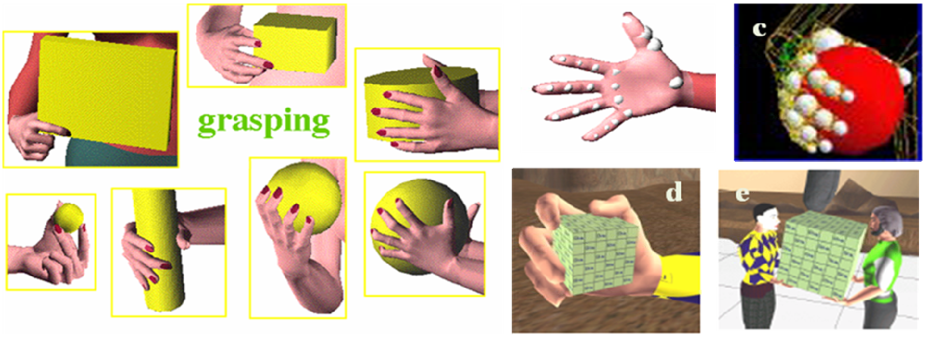


Fig. 1.13. a) grasping taxonomy b-c) multi-sensors d) one-hand grasping e) multiple grasping

Another approach used in the Jack system [66,67] that is based on Cutkosy's grasp classification [63] is described in [68]. This approach uses specialized controllers for each different type of grasp to close the fingers. It uses Parallel Transition Networks (PaT-Nets) [69] to control each finger. Transitions in the PaT-Nets are triggered by collisions between the fingers and the object. Different PaT-Nets are used to simulate different grasp types, the differing responses of the different PaT-Nets actually define how the grasp is to take place.

Abaci et al. [70] introduces a new grasping framework, which brings together a tubular feature classification algorithm, a hand grasp posture generation algorithm and an animation framework for human-object interactions. This unique combination is capable of handling grasping tasks within the proper context of virtual human object manipulation. This is very important since how an object is to be grasped depends strongly on how it is be used. The method has the advantage that it can work with relatively complex objects, where manual approximation with simple geometrical primitives may not be possible or practical. Furthermore, the method supports many intuitive parameters for controlling the grasping posture, such as the finger spread or the thumb configuration. Since the grasp parameters are specified as ranges, it is possible to generate a different posture each time a virtual human attempts to grasp an object, depending on the current configuration of the virtual human.

The algorithm to detect tubular features is called Plumber and it is a specialized shape classification method for triangle meshes. The Plumber method analyses the shape of an object by studying how the intersection of spheres centered at the mesh vertices evolve while the sphere radius changes. For example, for a thin limb, the curve of intersection between the mesh and a sphere will be simply connected for a small radius and then will rapidly split into two components when the radius increases and becomes greater than the tube size. A detailed description of the shape analysis technique which uses intersecting sphere and of the Plumber method can be found in [71, 72]. The algorithm segments a surface into connected components that are either body parts or elongated features, that is, handle-like and protrusion-like features, together with their concave counterparts, i.e. narrow tunnels and wells. The segmentation can be done at single or multi-scale, and produces a shape graph which codes how the tubular components are attached to the main body parts. Moreover, each tubular feature is represented by its skeletal line and an average cross-section radius.

Grasping is perhaps the most important part of a manipulation sequence, but it is not alone. A full sequence can consist of walking and reaching to the object, looking at it, grasping it multiple times, and keeping the hands constrained to the object while it is moving. Therefore, the smart objects are required to provide a full manipulation sequence, putting the grasping action into the proper context.

1.3.11 Motion Planning

An issue commonly encountered in Virtual Character animation is the problem of collisions with the environment. This problem is common to locomotion and object manipulation. Most animation algorithms (e.g. inverse kinematics) operate only on the Virtual Character and do not take the environment into account. When these motions are played, collisions between the virtual human and the scene may occur, detracting from the believability of the virtual environment. If care is taken during the design stage, the probability of collisions happening can be reduced, however, it is not possible to completely eliminate these, especially if we are not able to directly control what is happening in the virtual environment (e.g. if Virtual Characters are present).

In the field of robotics, researchers have been working on motion planning methods for robots to avoid collisions [73]. These methods can be applied to Virtual Characters, but a virtual character is an articulated structure with many degrees of freedom, therefore the dimensionality of the search space is very high. Methods based on probabilistic roadmaps [74, 75] are particularly suitable for structures of this complexity. A probabilistic roadmap is a data structure (graph) that is used to capture the connectivity of the search space. Nodes in the roadmap correspond to randomly–sampled configurations of the robot (e.g. joint angles) and an edge between two nodes in the roadmap means that the robot is able to move between corresponding configurations, by means of a local planner. Among these, Visibility–based Roadmap [76] construction techniques aim at reducing the number of nodes while the Rapidly–exploring Random Trees (RRT) [75, 77] focus on sufficiently exploring the configuration space at the expense of increasing the number of nodes.

The latest trend in motion planning for Virtual Characters is the use of motion capture data together with roadmap techniques. In [78], the authors attempt to solve the problem of biped locomotion by using randomly–sampled feet positions to construct the roadmap, which is augmented afterwards with a posture transition graph. Nodes in the roadmap are connected using data from input motion clips. In [79], motion planning algorithms based on probabilistic roadmaps are used to control 22 degrees of freedom (DOFs) of human-like characters in interactive applications. The main purpose is the automatic synthesis of collision-free reaching motions for both arms, with automatic column control and leg flexion. Generated motions are collision-free, in equilibrium, and respect articulation range limits. In order to deal with the high (22) dimension of the configuration space, the random distribution of configurations is biased to favor postures most useful for reaching and grasping. Figure 14 shows examples. In addition, there are extensions in order to interactively generate object manipulation sequences: a probabilistic inverse kinematics solver for proposing goal postures matching pre-designed grasps; dynamic update of roadmaps when obstacles change position; online planning of object location transfer; and an automatic stepping control to enlarge the character's reachable space. The work in [80] also focuses on the problem of object manipulation. The path of the object to be moved is computed using the RRT algorithm. An inverse kinematics algorithm generates poses that match the object position and orientation. Using soft constraints, it also biases the poses towards those in a posture database. As commonly seen with many virtual humans motion planning methods, some post–processing steps are used to increase the realism of the generated motions. The authors aim to reduce the dimensionality of the configuration space by planning only for the position and orientation of the object being manipulated.

Most existing work targeting motion planning for virtual humans assume that the virtual environment is static. However, if motion planning is to be used for object manipulation, then it is important to consider dynamics of the environment. Possible changes in the workspace can be included as additional dimensions in the configuration space, but a large number of dimensions is undesirable since it will reduce planning performance. To avoid this, dynamic roadmaps (DRM) can be used. For

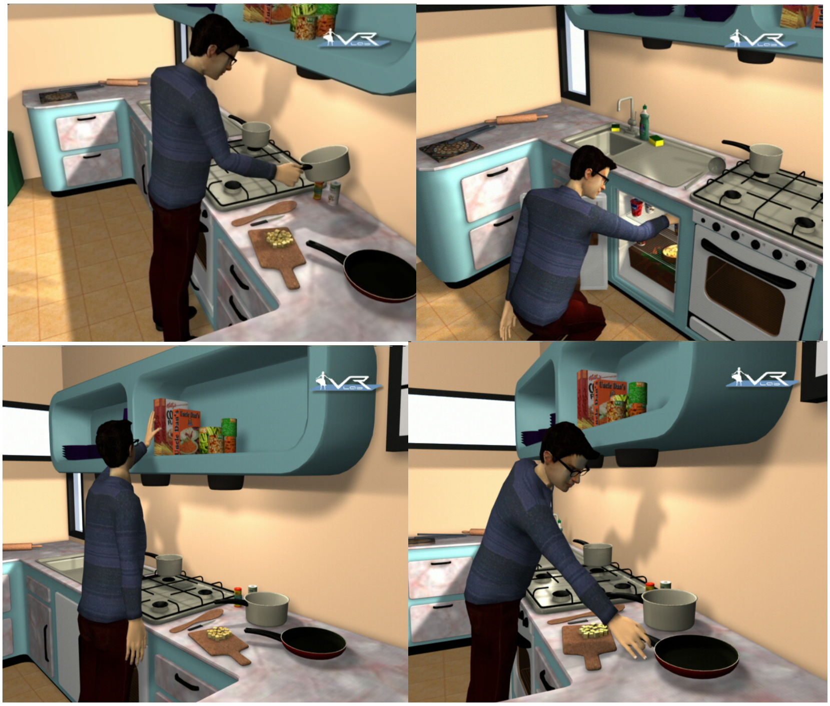


Fig. 1.14. Reaching examples

example, the work of Kallmann in [81] proposes a method to construct a dynamic roadmap [82] on top of a RRT planner, for application to a humanoid robot.

1.4 Facial Animation

The goal of facial animation systems has always been towards obtaining a high degree of realism using optimum resolution facial mesh models and effective deformation techniques. Various muscle based facial models with appropriate parameterized animation systems have been effectively developed for facial animation [Parke 1982][Waters 1987][Terzopoulos et al. 1990]. The Facial Action Coding System [Friesen 1978] defines high-level parameters for facial animation, on which several other systems are based. Most facial animation systems typically follow the following steps:

- Define an animation structure on a facial model by parameterization.
- Define "building blocks" or basic units of the animation in terms of these parameters, e.g. static expressions and visemes (visual counterparts of phonemes).
- Use these building blocks as key frames and define various interpolation and blending functions on the parameters to generate words and sentences from visemes and emotions (See Fig. 1.15) from expressions. The interpolation and blending functions contribute to the realism for a desired animation effect.
- Generate the mesh animation from the interpolated or blended key-frames. Given the tools of parameterized face modeling and deformation, the most challenging task in facial animation is the design of realistic facial expressions and visemes.

The complexity of the key-frame based facial animation system increases when we incorporate natural effects such as co-articulation for speech animation and blending between a variety of facial expressions during speech. The use of speech synthesis systems and the subsequent application of co-articulation to the available temporized phoneme information is a widely accepted approach [Grandstome 1999][Hill et al. 1988]. Coarticulation is a phenomenon observed during fluent speech, in which facial movements corresponding to one phonetic or visemic segment are influenced by those corresponding to the neighboring segments. Two main approaches taken for co-articulation are by Pelachaud [1991] and Cohen and Massaro [1993]. Both these approaches have been based on the classification of phoneme groups and their observed interaction during speech pronunciation. Pelachaud arranged the phoneme groups according to the deformability and context dependence in order to decide the influence of the visemes on each other. Muscle contraction and relaxation times were also considered and the Facial Action Units were controlled accordingly.

For Facial animation, the MPEG-4 standard is particularly important [MPEG-4]. The Facial Definition Parameter (FDP) set and the Facial Animation Parameter (FAP) set are designed to encode facial shape, as well as animation of faces thus reproducing

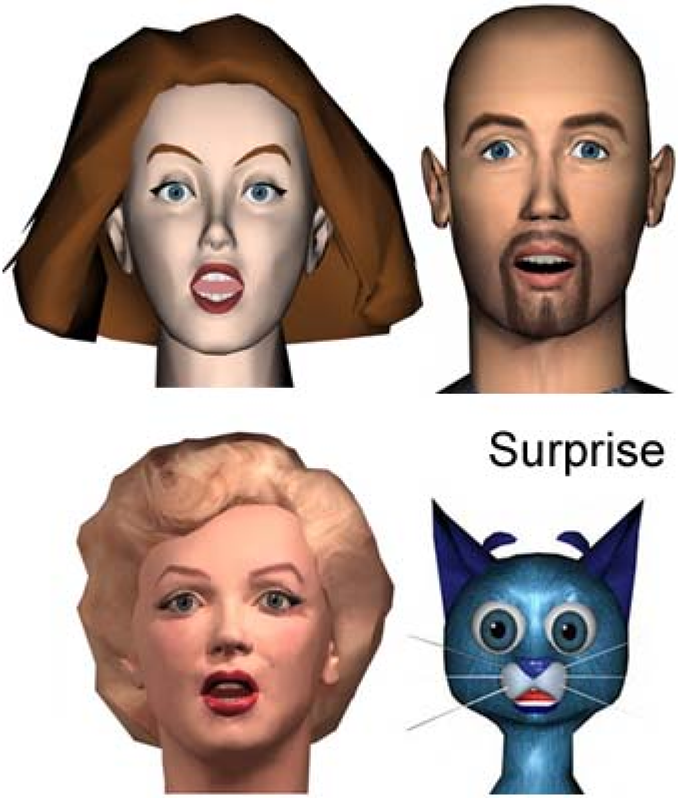


Fig. 1.15. Facial expressions

expressions, emotions and speech pronunciation. The FDPs are defined by the locations of the feature points and are used to customize a given face model to a particular face. Figure 1.16 shows the FDPs. They contain 3D feature points such as mouth corners and contours, eye corners, eyebrow centers, etc. FAPs are based on the study of minimal facial actions and are closely related to muscle actions. Each FAP value is simply the displacement of a particular feature point from its neutral position expressed in terms of the Facial Animation Parameter Units (FAPU). The FAPUs correspond to fractions of distances between key facial features (e.g. the distance between the eyes). For exemple, the MPEG-4-based facial animation engine [83] for animating 3D facial models works in real timeand is capable of displaying a variety of facial expressions, including speech pronunciation with the help of 66 low level Facial Animation Parameters (FAPs).

Recently, the efforts in the field of phoneme extraction have resulted into software systems capable of extracting phonemes from synthetic as well as natural speech and generating lip synchronized speech animation from these phonemes thus creating a complete talking head system. It is possible to mix emotions with speech in a natural way, thus imparting the virtual character an emotional behavior. The ongoing efforts

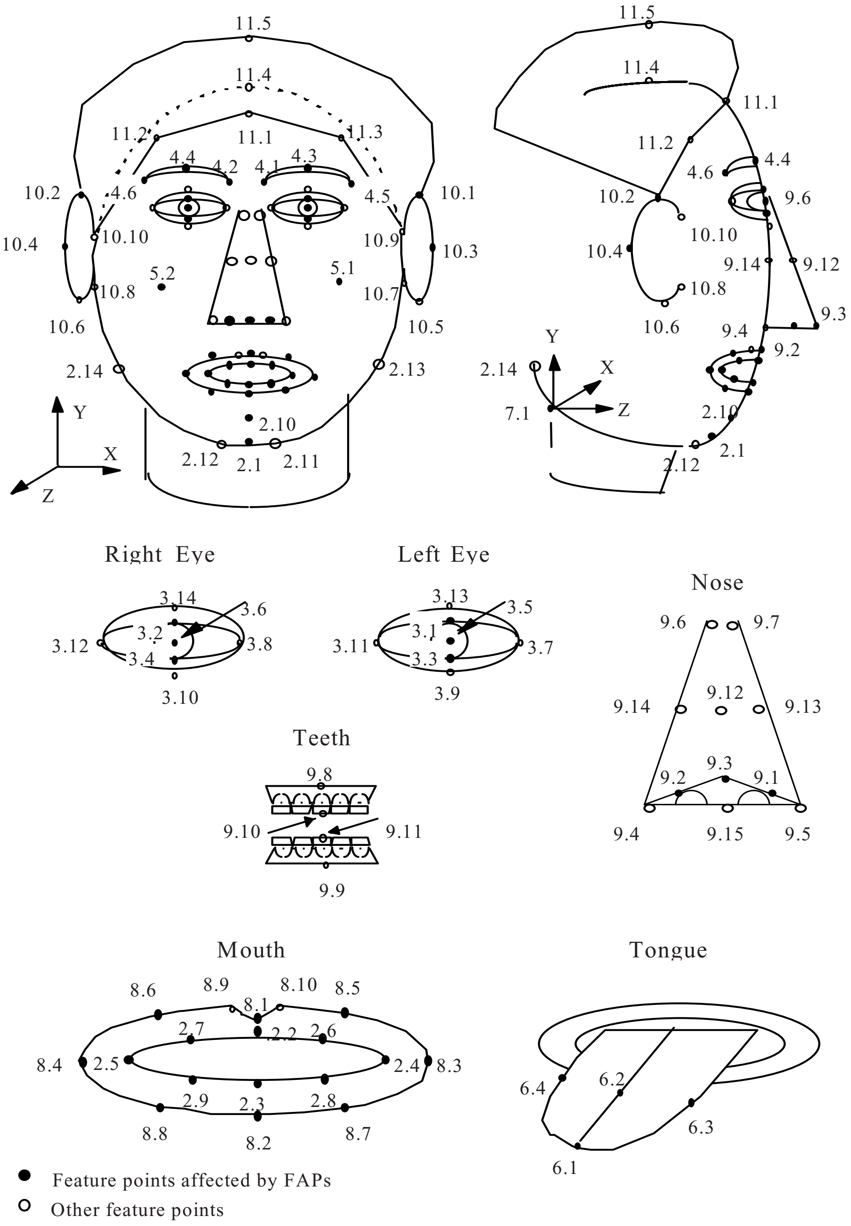


Fig. 1.16. FDP Feature Points

are concentrated on imparting "emotional" autonomy to the virtual face enabling a dialogue between the real and the virtual humans with natural emotional responses. Kshirsagar and Magnenat-Thalmann [2001] use a statistical analysis of the facial feature point movements. As the data is captured for fluent speech, the analysis reflects the dynamics of the facial movements related to speech production. The results of the analysis were successfully applied for a more realistic speech animation. Also, this has enabled us to easily blend between various facial expressions and speech. Use of MPEG-4 feature points for data capture and facial animation enabled us to restrict the quantity of data being processed, at the same time offering more flexibility with respect to the facial model. We would like to further improve the effectiveness of the expressive speech by use of various time envelopes for the expressions that may be linked to the meaning of the sentence. Kshirsagar and Magnenat-Thalmann [2002] have also developed a system incorporating a personality model for an emotional autonomous virtual human.

1.5 Autonomous Characters

1.5.1 Why Autonomous Virtual Characters?

Virtual Characters are not just for movies and games. In the future, they can be at the heart of the simulation of activities for a variety of purposes, including education and training, treatment of psychological problems, and emergency preparedness. Imagine the following scenarios:

- A user is being trained to perform some complex task, such as repairing a copy machine. He uses an interactive user manual, where an autonomous character plays an expert, showing him how to proceed. At every stage, the user is able to see what to do next, even when mistakes are made.
- A therapist is helping a patient overcome a fear of public speaking. To overcome this fear, the patient has to perform while immersed in a virtual environment consisting of a seminar room and a virtual audience, which can react to the user in an autonomous way. The therapist can choose the type of virtual audience (for instance, one that is aggressive or sexist) that will result in a more effective treatment for the patient.

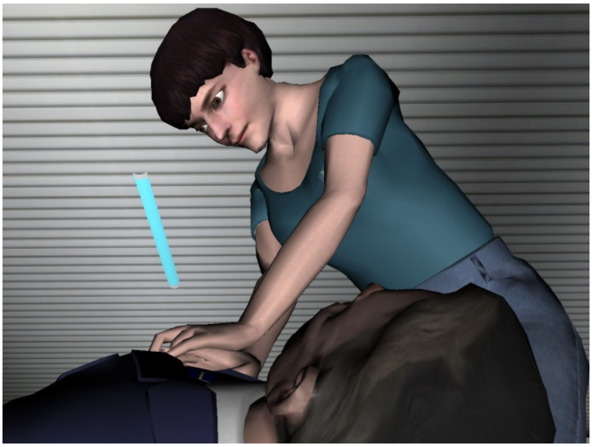


Fig. 1.17. Virtual Assistant for Basic Life Support

- A user is learning basic life support (BLS) procedures. She is immersed in a virtual setting, and discovers a victim lying on the ground. She has to give him BLS through her proxy, a virtual assistant. The user navigates the scene, assesses the situation, and makes decisions by issuing natural voice commands. The Virtual Assistant waits for commands and executes the actions. If the user's commands are correct, the victim recovers. In cases where the user provides incorrect commands, the Virtual Assistant may refuse to do harm to the victim; in such situations, the Virtual Assistant may prompt the user for retrial, or may suggest an alternative possibility. Fig. 1.17 shows the scene.

To be able to accomplish such tasks, Virtual Characters should be able to act on their own. This means that they should be Autonomous Virtual Characters (AVCs).

1.5.2 Properties of Autonomous Virtual Characters

Autonomy is generally the quality or state of being self-governing. Rather than acting from a script, an AVC is aware of its changing virtual environment, making its own decisions in real time in a coherent and effective way. An AVC should appear to be spontaneous and unpredictable, making the audience believe that the character is really alive and has its own will.

To be autonomous, an AVC must be able to perceive its environment and decide what to do to reach an intended goal. The decisions are then transformed into motor control actions, which are animated so that the behavior is believable. Therefore, an AVC's behavior consists of always repeating the following sequence: perception of the environment, action selection, and reaction.

The problem with designing AVCs is determining how to decide on the appropriate actions at each point in time, to work toward the satisfaction of the current goal, which represents the AVC's most urgent need. At the same time, there is a need to pay attention to the demands and opportunities coming from the environment, without neglecting, in the long term, the satisfaction of other active needs.

There are four properties that determine how AVCs make their decisions: perception, adaptation and intelligence, memory, and emotions.

Perception. Perception of the elements in the environment is essential for AVCs, as it gives them an awareness of what is changing. An AVC continuously modifies its environment, which, in turn, influences its perceptions. Therefore, sensorial information drastically influences AVC behavior. This means that we cannot build believable AVCs without considering the way they perceive the world and each other. It is tempting to simulate perception by directly retrieving the location of each perceived object straight from the environment. To realize believable perception, AVCs should have sensors that simulate the functionality of their organic counterparts, mainly for vision, audition, and tactile sensation. These sensors should be used as a basis for implementing everyday behaviors, such as visually directed locomotion, responses to sounds and utterances, and the handling of objects. What is important is the functionality of a sensor and how it filters the information flow from the environment. It is not necessary or efficient to model sensors with biological accuracy. Therefore, virtual eyes may be represented by a Z-buffered color image representing a character's vision [84]. Each pixel of the vision input has the semantic information giving the object projected on this pixel, and numerical information giving the distance to this object. So, it is easy to know, for example, that there is a table just in front at 3 meters. A virtual nose, or other tactile point-like sensors, may be represented by a simple function evaluating the global force field at the sensor's location. The virtual ear of a character may be represented by a function returning the ongoing sound events. With these virtual sensors, AVCs should be able to perceive the virtual world in a way that is very similar to the way they would perceive the real one. In a typical behavioral animation scene, the actor perceives the objects and the other actors in the environment, which provides information on their nature and position. This information is used by the behavioral model to decide the action to take, which results in a motion procedure. The synthetic actor perceives his environment from a small window in which the environment is rendered from his point of view..

Adaptation and intelligence. Adaptation and intelligence define how the character is capable of reasoning about what it perceives, especially when unpredictable events happen. An AVC should constantly choose the best action so that it can survive in its environment and accomplish its goals. As the environment changes, the AVC should be able to react dynamically to new elements, so its beliefs and goals may evolve over time. An AVC determines its next action by reasoning about what it knows to be true at a specific time. Its knowledge is decomposed into its beliefs and internal states, goals, and plans, which specify a sequence of actions required to achieve a specific goal. When simulating large groups or communities of AVCs, it is possible to use bottom-up solutions that use artificial life techniques, rather than top-down, plan-based approaches, such as those that are common in artificial intelligence. This allows new, unplanned behaviors to emerge.

Memory. It is necessary for an AVC to have a memory so that similar behaviors can be selected when predictable elements reappear. Memory plays an important role in the modelling of autonomy, as actions are often decided based on memories. But imagine an AVC in a room containing 100 different objects. Which objects can be considered memorized by the virtual character? It is tempting to decide that whenever an object is seen by the AVC, it should be stored in its memory. But if you consider humans, nobody is able to remember every single object in a room. Therefore, the memory of a realistic AVC should not be perfect either. Noser et al. [85] proposed the use of an octree as the internal representation of the environment seen by an actor because it can represent the visual memory of an actor in a 3D environment with static and dynamic objects.

Emotions. The believability of an AVC is made possible by the emergence of emotions clearly expressed at the right moment. An emotion is an emotive reaction to a perception that induces a character to assume a physical response, facial expression, or gesture, or select a specific behavior. The apparent emotions of an AVC and the way it reacts are what give it the appearance of a living being with needs and desires. Without them, an actor would just look like an automaton. Apart from making them appear more realistic, AVCs' visible emotions can provide designers with a direct way of influencing the user's emotional state. To allow AVCs to respond emotionally to a situation, they could be equipped with a computational model of emotional behavior. Emotionally related behavior, such as facial expressions and posture, can be coupled with this computational model, which can be used to influence their actions. The development of a good computational model is a challenge.

Motivations. Motivation is also a key cause of action and we will consider basic motivations as essential to model Life in the Virtual World, providing a true Virtual Life. We adapt basic actor motivations and needs to urban situations, where most actions involve interactions with the environment, especially manipulation of objects in natural situations like eating or drinking. We focus on common-life situations, where the actor senses and explores his environment, and following an action selection mechanism, determines the suitable actions to take. For this, we can consider a mechanism of action selection based on a free flow hierarchy associated to a hierarchical classifier [86]. The hierarchy of our model will contain four levels (depicted in Figure 1.18):

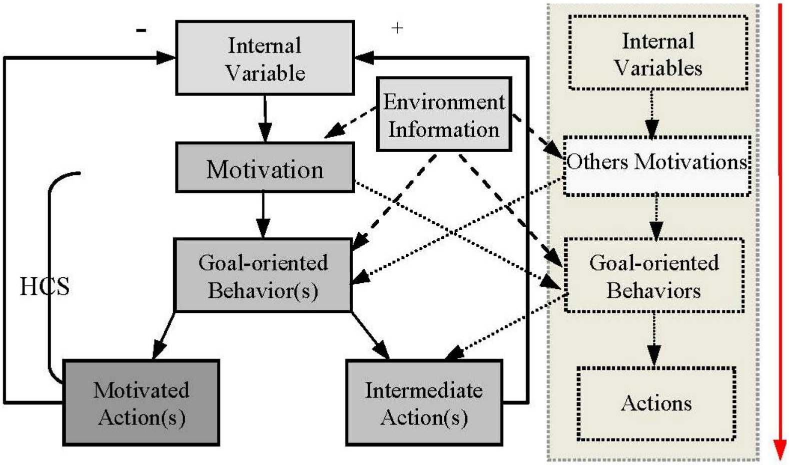


Fig. 1.18. Simplified motivational model of action selection for virtual humans

The free flow hierarchy [87] permits to take into account different types of motivations and also information coming from the environment perception. The key idea is that, during the propagation of the activity in the hierarchy, no choices are made before the lowest level in the hierarchy represented by the actions is reached. The hierarchical classifier [88] provides a good solution to model complex hierarchies by reducing the search domain of the problem, and using rules with weights. Also, we can easily make behavioural sequences (composed by a sequence of actions). As a result, the virtual character can move to a specific place to perform a physiological action and satisfy the motivations no matter where it is. In another words, the main role of the action selection mechanism is to maintain the internal variables under the threshold by choosing the correct actions. Actions involving interactions are preferably chosen because they are defined to be directly beneficial for the virtual human. Otherwise, the virtual human is instructed to walk and reach the place where the motivation can be satisfied. During the simulation, the model is fed with parameters describing the current state of the actor concerning each of the motivations, and by flowing inside the hierarchical structure, will correctly trigger the concerned actions. After an action is selected as a response to satisfy one of the actor's motivations, the state parameter of the internal variable is adapted accordingly. An example is shown in Figure 1.19.

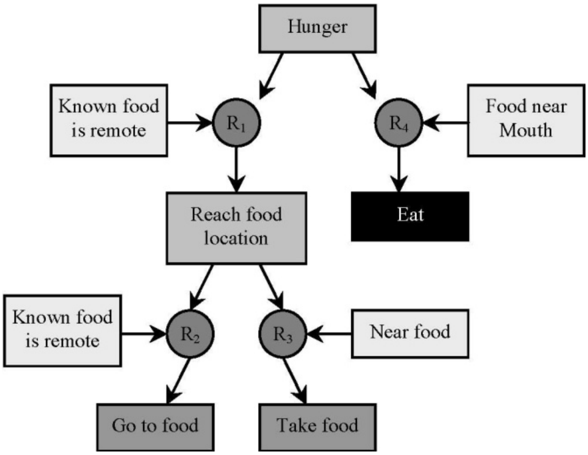


Fig. 1.19. Virtual life simulation: by default, the Virtual Human is working, waiting for a motivation (for instance, drinking) to be activated. The food is kept into the kitchen, as seen in the background.

1.6 Crowd Simulation

Real-time crowds bring different challenges compared to the systems either involving small number of interacting characters (for example, the majority of contemporary computer games), or non-real-time applications (as crowds in movies, or visualizations of crowd evacuations after off-line model computations). In comparison with single-agent simulations, the main conceptual difference is the need for efficient variety management at every level, whether it is visualization, motion control, animation or sound rendering. As everyday experiences hint, virtual humans composing a crowd should look different, move different, react different, sound different and so forth. Even if assuming perfect simulation of a single virtual human would be possible, still creating a simulation involving multiple such humans would be a difficult and tedious task. Methods easing control of many characters are needed; however such methods should still preserve ability to control individual agents. For a extended study of crowd simulation, the reader should see [89].

In comparison with non-real-time simulations, the main technical challenge is increased demand on computational resources whether it is general processing power, graphics performance or memory space. One of the foremost constraining factors for real-time crowd simulations is crowd rendering. Fast and scalable methods both to compute behavior, able to take into account inputs not known in advance, and to render large and varied crowds, are needed. While non-real-time simulations are able to take advantage of knowing a full run of the simulated scenario (and therefore, for example, can run iteratively over several possible options selecting the globally best solution), real-time simulations have to react to the situation as it unfolds in the moment.

Animating crowds [90] is challenging both in character animation and a virtual city modeling. Though different textures and colors may be used, the similarity of the virtual people would be soon detected by even non-experts, say, "everybody walks the same in this virtual city!" . It is, hence, useful to have a fast and intuitive way of generating motions with different personalities depending on gender, age, emotions, etc., from an example motion, say, a genuine walking motion. The problem is basically to be able to generate variety among a finite set of motion requests and then to apply it to either an individual or a member of a crowd. It also needs very good tools to tune the motion [91].

Bouvier et al. [92],[93] used combination of particle systems and transition networks to model human crowds in visualization of urban spaces. Lower level enabled people to avoid obstacles using attraction and repulsion forces analogous to physical electric forces. Higher level behavior is modeled by transition networks with transitions depending on timing, visiting of certain points, changes of local densities and global events. Brogan and Hodgins [94] simulated group behavior for systems with significant dynamics. They presented algorithm for controlling the movements of creatures traveling as a group. Musse and Thalmann's [90] proposed solution addresses two main issues: i) crowd structure and ii) crowd behavior. Considering crowd structure, their approach deals with a hierarchy composed of crowd, groups and agents, where the groups are the most complex structure containing the information to be distributed among the individuals. Concerning crowd behavior, the virtual agents are endowed with different levels of autonomy. For emergent crowds, Ulicny and Thalmann [95] proposed a behavior model based on combination of rules [Rosenblum et al. 1998] and finite state machines [96] for controlling agent's behavior using layered approach. First layer deals with the selection of higher-level complex behavior appropriate to agent's situation, second layer implements these behaviors using low-level actions provided by the virtual human [97]. At the higher level, rules select complex behaviors (such as flee) according to agent's state (constituted by attributes) and the state of the virtual environment (conveyed by events).

In terms of rendering, the goal of a real-time crowd visualizer [98] is to render a large number of entities according to the current simulation state, which provides the position, orientation, and velocity for each individual. System constraints are believability, real-time updates (25 frames per second) and a number of digital actors ranging in the tens of thousands. Also actors are made believable by varying their appearance (textures and colors) and animation; we may also add accessories like hats, glasses, or mobile phones (see Figure 1.20). Their graphical representation is derived from a template, which holds all the possible variations. Thus, with only a limited set of such templates, we can achieve a varied crowd, leading to considerable time savings for designers.

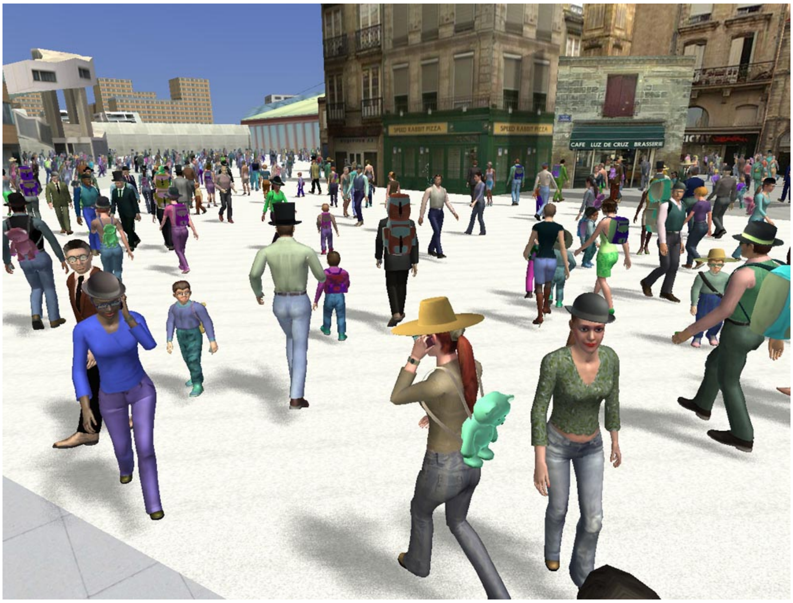


Fig. 1.20. Crowd simulation with varieties, accessories; the rendering is based on three fidelities

Each human in the visualization system is called an instance and is derived from a template. Individualization comes from assigning a specific gray scale texture and a color combination for each identifiable region. Instances have individualized walk velocities and are animated by blending the available walk animations.

The rendering pipeline advances consecutively in four steps. The first one consists in culling, that is, determining visibility, and choosing the rendering fidelity for each simulated human. Three fidelities are defines: dynamic meshes, static meshes, and impostors. The next step of the pipeline is the rendering of dynamic meshes, which are the most detailed fidelity capable to interpolate animations based on skeletal postures. According to the current instance state (linear and angular walk velocities and time), animations are retrieved from the database and interpolated, yielding a smooth animation, with continuous variations of velocities, and no foot-sliding. The resulting skeletal posture is sent to a hardware vertex shader and fragment shader deforming and rendering the human on the graphics card. Then, static meshes (also called baked or predeformed) constitute the second rendering fidelity, which keeps a pre-transformed set of animations using the lowest resolution mesh of the deformed mesh in the previous step. Pre-computing deformations allows substantial gains in speed, but constrains the animation variety and smoothness. The final rendering fidelity is the billboard model which, compared to previous approaches, uses a simplified scheme of sampling and lighting. Worldaligned billboards are used, with the assumption that the camera will never hover directly above the crowd. Thus, only sampled images around the waist level of the character are needed. E.g. the templates are sampled at 20 different angles, for each of the 25 keyframes composing a walk animation. When constructing the resulting texture, the bounding box of each sampled frame is detected to pack them tightly together. When rendering billboarded pedestrians, cylindrical lighting is applied: each vertex normal is set to point in the positive Z direction, plus a small offset on the X axis, so that it points slightly outside the frame. We then interpolate the light intensity for each pixel in the fragment shader.

Planning crowd motion in real time is a very expensive task, which is often decoupled into two distinct parts: path planning and obstacle avoidance. Path planning consists in finding the best way to reach a goal. Obstacles can either be other pedestrians or objects that compose the environment. The path selection criteria are the avoidance of congested zones, and minimization of distance and travel time. Path planning must also offer a variety of paths to spread pedestrians in the whole scene. Avoidance, on the other hand, must inhibit collisions of pedestrians with obstacles. For real-time simulations, such methods need to be efficient as well as believable. Multiple motion planning approaches for crowds have been introduced. As of today, several fast path planning solutions exist. Avoidance however, remains a very expensive task. Agent-based methods offer realistic pedestrian motion planning, especially when coupled with global navigation. This approach gives the possibility to add individual and cognitive behaviors for each agent, but becomes too expensive for a large number of pedestrians. Potential field approaches [99] handle long and short-term avoidance. Long term avoidance predicts possible collisions and inhibits them. Short term avoidance intervenes when long-term avoidance alone cannot prevent collisions. These methods offer less believable results than agent-based approaches, because they do not provide the possibility to individualize each pedestrian. However, this characteristic also implies much lower computational costs.

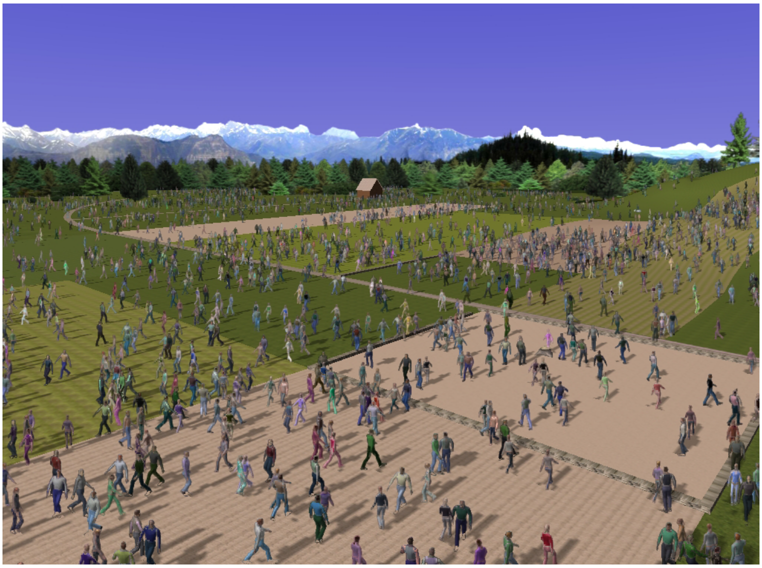


Fig. 1.21. Pedestrians using an hybrid motion planning architecture to reach their goal and avoid each other

Recently, an hybrid architecture [100] has been proposed to handle realistic crowd motion planning in real time. In order to obtain high performance, the approach is scalable. The authors divide the scene into multiple regions of varying interest, defined at initialization and modifiable at runtime. According to its level of interest, each region is ruled by a different motion planning algorithm. Zones that attract the attention of the user exploit accurate methods, while computation time is saved by applying less expensive algorithms in other regions. The architecture also ensures that no visible disturbance is generated when switching from an algorithm to another. The results shows that it is possible to simulate up to ten thousand pedestrians in real time with a large variety of goals. Moreover, the possibility to introduce and interactively modify the regions of interest in a scene offers a way for the user to select the desired performance and to distribute the computation time accordingly. A simulation of pedestrians taking advantage of this architecture is illustrated in Figure 1.21.

References

1. Fetter, W.A.: A Progression of Human Figures Simulated by Computer Graphics. IEEE Computer Graphics and Applications 2(9), 9–13 (1982)
2. Muybridge, E.: The Human Figure in Motion. Dover Publications. New York City (1955)
3. Blakeley, F.M.: CYBERMAN, Chrysler Corporation, Detroit, Michigan (June 1980)
4. Evans, S.M.: User's Guide for the Program of Combiman, Report AMRLTR-76-117, University of Dayton, Ohio (1976)
5. Dooley, M.: Anthropometric Modeling Programs – A Survey. IEEE Computer Graphics and Applications 2(9), 17–25 (1982)
6. Bonney, M.C., Case, K., Hughes, B.J., Schofield, N.A., Williams, R.W.: Computer Aided Workplace Design using SAMMIE. In: Proc. Ergonomics Research Society Annual Conference, Cardiff (1972)
7. Parke, F.I.: Computer Generated Animation of Faces. In: Proc. ACM annual conference (1972)
8. Poter, T.E., Willmert, K.D.: Three-Dimensional Human Display Model. Computer Graphics 9(1), 102–110 (1975)
9. Herbison-Evans, D.: Animation of the Human Figure, Technical Report CS-86-50, University of Waterloo Computer Science Department (November 1986)
10. Badler, N.I., Smoliar, S.W.: Digital Representations of Human Movement. Computing Surveys 11(1), 19–38 (1979)
11. Calvert, T.W., Patla, A.: Aspects of the Kinematic Simulation of Human Movement. IEEE Computer Graphics and Applications 2(9), 41–50 (1982)
12. Molet, T., Aubel, A., Çapin, T., Carion, S., Lee, E., Magnenat Thalmann, N., Noser, H., Pandzic, I., Sannier, G., Thalmann, D.: Anyone for Tennis? Presence 8(2), 140–156 (1999)
13. Magnenat-Thalmann, N., Laperrière, R., Thalmann, D.: Joint-dependent local deformations for hand animation and object grasping In: Proceedings on Graphics interface 1988, pp. 26–33. Canadian Information Processing Society (1988)
14. Lasseter, J.: Principles of traditional animation applied to 3D computer animation. In: SIGGRAPH 1987: Proceedings of the 14th annual conference on Computer graphics and interactive techniques, pp. 35–44. ACM Press, New York (1987)
15. Chadwick, J.E., Haumann, D.R., Parent, R.E.: Layered construction for deformable animated characters. In: SIGGRAPH 1989: Proceedings of the 16th annual conference on Computer graphics and interactive techniques, pp. 243–252. ACM Press, New York (1989)

16. Denavit, J., Hertenberg, R.S.: Kinematics notation for lower pair mechanism based on matrices. Appl. mech. 77, 215–221 (1955)
17. Steindler, A.: Kinesiology of the Human Body. Charles C. Thomas Publisher, Springfield Illinois 1955 (1955)
18. Azuola, F., Badler, N., Ho, P., Kakadiaris, I., Metaxas, D., Ting, B.: Building anthropometry-based virtual human models. In: Proc. IMAGE VII Conference (1994)
19. Shen, J., Thalmann, N.M., Thalmann, D.: Human Skin Deformation from Cross-Sections Computer Graphics Int. 1994 (1994)
20. Shen, J., Thalmann, D.: Interactive Shape Design Using Metaballs and Splines. In: Implicit Surfaces 1995, pp. 187–196 (1995)
21. Wilhelms, J.: Modeling Animals with Bones, Muscles, and Skin, University of California at Santa Cruz (1995)
22. Wilhelms, J., Gelder, A.V.: Anatomically based modeling. In: SIGGRAPH 1997: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pp. 173–180. ACM Press/Addison-Wesley Publishing Co. (1997)
23. Scheepers, F., Parent, R.E., Carlson, W.E., May, S.F.: Anatomy-based modeling of the human musculature. In: SIGGRAPH 1997: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pp. 163–172. ACM Press/Addison-Wesley Publishing Co. (1997)
24. Lee, W., Gu, J., Magnenat-Thalmann, N., Gross, M., Hopgood, F.R.A. (ed.): Generating Animatable 3D Virtual Humans from Photographs. In: Computer Graphics Forum (Eurographics 2000), vol. 19(3) (2000)
25. Allen, B., Curless, B., Popovic, Z.: Articulated body deformation from range scan data. ACM Trans. Graph. 21, 612–619 (2002)
26. Seo, H., Magnenat-Thalmann, N.: An example-based approach to human body manipulation Graph. Models, vol. 66, pp. 1–23. Academic Press Professional, Inc., London (2004)
27. Magnenat Thalmann, N., Thalmann, D.: Complex Models for Animating Synthetic Actors. IEEE Computer Graphics and Applications 11(5), 32–44 (1991)
28. Parke, F.: Parameterized Models for Facial Animation. IEEE Computer Graphics and Applications 2(9), 61–68 (1982)
29. Kochanek, D.H., Bartels, R.H.: Interpolating Splines with Local Tension, Continuity, and Bias Control. In: Proc. SIGGRAPH 1984, pp. 33–41 (1984)
30. Fu, K.S., Gonzalez, Lee, C.S.G.: Robotics, Control, Sensing, Vision, and Intelligence, pp. 52–76, 84–102, 111–112. McGraw-Hill, Inc., New York (1987)
31. Witkin, A., Popovic, Z.: Motion warping. In: Proc. SIGGRAPH 1995, pp. 105–108 (1995)
32. Popovic, Z., Witkin, A.: Physically based motion transformation. In: Proc. SIGGRAPH 1999, pp. 11–20 (1999)
33. Bruderlin, A., Williams, L.: Motion signal processing. In: Proc. SIGGRAPH 1995, pp. 97–104 (1995)
34. Gleicher, M.: Retargeting motion to new characters. In: Proc. SIGGRAPH 1998, pp. 33–42 (1998)
35. Monzani, J.-S., Baerlocher, P., Boulic, R., Thalmann, D.: Using an Intermediate Skeleton and Inverse Kinematics for Motion Retargeting. In: Proc. Eurographics 2000, pp. 11–19 (2000)
36. Bindiganavale, R., Badler, N.I.: Motion abstraction and mapping with spatial constraints. In: Magnenat-Thalmann, N., Thalmann, D. (eds.) Modeling and Motion Capture Techniques for Virtual Environments, Geneva, Switzerland, November 1998. LNCS (LNAI), pp. 70–82. Springer, Heidelberg (1998)

37. Jehee, L., Shin, S.Y.: A hierarchical approach. In: Proc. SIGGRAPH 1999, pp. 39–48 (1999)
38. Lim, I.S., Thalmann, D.: Solve Customers Problems: Interactive Evolution for Tinkering with Computer Animation. In: Proc. 2000 ACM Symposium on Applied Computing (SAC 2000), pp. 404–407 (2000)
39. Huston, R.L.: Multibody dynamics. Butterworth-Heinemann, Stoneham (1990)
40. Featherstone, R.: Robot Dynamics Algorithms. Kluwer Academic Publishers, Dordrecht (1986)
41. Mac Kenna, M., Zeltzer, D.: Dynamic Simulation of a Complex Human Figure Model with Low Level Behavior Control. Presence 5(4), 431–456 (1996)
42. Boulic, R., Magnenat-Thalmann, N., Thalmann, D.: A Global Human Walking Model with Real-time Kinematics Personification. The Visual Computer 6(6), 344–358 (1990)
43. Thalmann, D., Raupp Musse, S.: Crowd Simulation, ch. 4. Springer, Heidelberg (2007)
44. Multon, F., France, L., Cani-Gascuel, M.P., Debunne, G.: Computer animation of human walking: a survey. The Journal of Visualization and Computer Animation 10(1), 39–54 (1999)
45. Maya: Alias systems corp.
46. van de Panne, M.: From Footprints to Animation. Computer Graphics Forum, 211–223 (1997)
47. Choi, M., Lee, J., Shin, S.: Planning Biped Locomotion using Motion Capture Data and Probabilistic Roadmaps. ACM Transactions on Graphics (2003)
48. Wooten, W.L., Hodgins, J.K.: Simulating leaping, tumbling, landing and balancing humans. In: Proc. IEEE International Conference on Robotics and Automation (2000)
49. Boulic, R., Ulciny, B., Thalmann, D.: Versatile walk engine. Journal of Game Development 1, 29–50 (2004)
50. Bruderlin, A., Calvert, T.: Knowledge-driven, interactive animation of human running. In: Proc. Graphics Interface 1996, pp. 213–221 (1996)
51. Sun, H.C., Metaxas, D.N.: Automating Gait Generation. In: Proc. SIGGRAPH 2001, Annual Conference Series (2001)
52. Chung, S., Hahn, J.K.: Animation of Human Walking in Virtual Environments. In: Proc. Computer Animation 1999 (1999)
53. Bruderlin, A., Williams, L.: Motion signal processing. In: Proc. SIGGRAPH 1995. Annual Conference Series, pp. 97–104 (1995)
54. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure. In: Proc. SIGGRAPH 1995. Annual Conference Series, pp. 91–96 (1995)
55. Rose, C., Cohen, M.F., Bodenheimer, B.: Verbs and adverbs: Multidimensional motion interpolation. IEEE Computer Graphics and Applications 18(5), 32–41 (1998)
56. Park, S.I., Shin, H.J., Shin, S.Y.: On-line locomotion generation based on motion blending. In: Proc. SIGGRAPH/Eurographics Symposium on Computer Animation (2002)
57. Kovar, L., Gleicher, M.: Flexible automatic motion blending with registration curves. In: Proc. ACM SIGGRAPH/Eurographics Symposium on Computer Animation 2003, pp. 214–224 (2003)
58. Jolliffe, I.T.: Principal Component Analysis. Springer series in statistics. Springer, Heidelberg (1986)
59. Alexa, M., Müller, W.: Representing animations by principal components. In: Proc. Eurographics 2000, vol. 19(3), pp. 411–426 (2000)
60. Lim, I.S., Thalmann, D.: Construction of animation models out of captured data. In: Proc. IEEE Conf. Multimedia and Expo. (August 2002)
61. H-anim, humanoid animation working group (2004), http://www.hanim.org

62. Alexa, M.: Linear combination of transformations. In: Proc. SIGGRAPH 2002, pp. 380–387 (2002)
63. Cutkosky, M.R.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Transactions on Robotics and Automation 5, 269–279 (1989)
64. Mas, R., Boulic, R., Thalmann, D.: Extended grasping behavior for autonomous human agents. In: AGENTS 1997: Proceedings of the first international conference on Autonomous agents, pp. 494–495. ACM Press, New York (1997)
65. Mas, R., Thalmann, D.: A hand control and automatic grasping system for synthetic actors. Computer Graphics Forum 13(3), 167–177 (1994)
66. Phillips, C.B., Badler, N.I.: Jack: a toolkit for manipulating articulated figures. In: Proc. UIST 1988, Proc. SIGGRAPH symposium on User Interface Software, pp. 221–229 (1988)
67. Badler, N.I., Phillips, C.B., Webber, B.L.: Simulating humans: computer graphics animation and control. Oxford University Press, Inc., New York (1993)
68. Douville, B., Levison, L., Badler, N.I.: Task level object grasping for simulated agents. Presence 4(5), 416–430 (1996)
69. Trias, T.S., Chopra, S., Reich, B.D., Moore, M.B., Badler, N.I., Webber, B.L., Geib, C.W.: Decision networks for integrating the behaviors of virtual agents and avatars. In: Proc. 1996 Virtual Reality Annual International Symposium (VRAIS 1996), p. 156 (1996)
70. Abaci, T., Ciger, J., Thalmann, D.: Planning with Smart Objects. In: Proc. International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG 2005, Czech Republic (February 2005)
71. Mortara, M., Patane, G., Spagnuolo, M., Falcidieno, B., Rossignac, J.: Blowing bubbles for the multi-scale analysis and decomposition of triangle meshes. Algorithmica, Special Issues on Shape Algorithms 38(2), 227–248 (2004)
72. Mortara, M., Patane, G., Spagnuolo, M., Falcidieno, B., Rossignac, J.: Plumber: a method for a multi-scale decomposition of 3d shapes into tubular primitives and bodies. In: Proc. Ninth ACM Symposium on Solid Modeling and Applications SM 2004, pp. 339–344 (2004)
73. Latombe, J.C.: Robot Motion Planning. Kluwer Academic Publishers, Norwell (1991)
74. Kavraki, L., Svestka, P., Latombe, J.C., Overmars, M.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. Technical report, Stanford University, Stanford, CA, USA (1994)
75. LaValle, S.M.: Rapidly-exploring random trees: A new tool for path planning. Technical Report 98-11, Dept. of Computer Science, Iowa State University (October 1998)
76. Simon, T., Laumond, J.P., Nissoux, C.: Visibility based probabilistic roadmaps for motion planning. Advanced Robotics Journal 14(2) (2000)
77. Kuffner Jr., J.J., LaValle, S.M.: RRT-connect: An efficient approach to single-query path planning. In: Proc. ICRA 2000, pp. 995–1001 (2000)
78. Choi, M.G., Lee, J., Shin, S.Y.: Planning biped locomotion using mo-tion capture data and probabilistic roadmaps. ACM Transactions on Graphics 22(2), 182–203 (2003)
79. Kallmann, M., Aubel, A., Abaci, T., Thalmann, D.: Planning Collision-Free Reaching Motions for Interactive Object Manipulation and Grasping. In: Proc. Eurographics 2003, September 2003, vol. 22(3) (2003)
80. Yamane, K., Kuffner, J.J., Hodgins, J.K.: Synthesizing animations of human manipulation tasks. ACM Transactions on Graphics 23(3), 532–539 (2004)
81. Kallmann, M., Mataric, M.: Motion planning using dynamic roadmaps. In: Proc. ICRA 2004, vol. 5, pp. 4399–4404 (2004)

82. Leven, P., Hutchinson, S.: Motion planning using dynamic roadmaps. In: Proc. Fourth International Workshop on the Algorithmic Foundations of Robotics (WAFR), pp. 363–376 (2000)
83. Kshirsagar, S., Garchery, S., Magnenat Thalmann, N.: Feature Point Based Mesh Deformation Applied to MPEG-4 Facial Animation. In: Proceedings Deform 2000 (2000)
84. Renault, O., Magnenat-Thalmann, N., Thalmann, D.: A Vision-Based Approach to Behavioural Animation. Journal of Visualization and Computer Animation 1(1), 18–21 (1990)
85. Noser, H., Renault, O., Thalmann, D., Magnenat Thalmann, N.: Navigation for Digital Actors based on Synthetic Vision, Memory and Learning. Computers and Graphics 19(1), 7–19
86. de Sevin, E., Thalmann, D.: A motivational Model of Action Selection for Virtual Humans. In: Proc. Computer Graphics International (CGI). IEEE Computer Society Press, New York (2005)
87. Tyrrell, T.: The use of hierarchies for action selection (1993)
88. Donnart, J.Y., Meyer, J.A.: Learning Reactive and Planning Rules in a Motivation-ally Autonomous Animal. IEEE Transactions on Systems, Man, and Cybernetics, part B: Cybernetics 26(3), 381–395 (1996)
89. Thalmann, D., Musse, S.R.: Crowd Simulation. Springer, Heidelberg (2007)
90. Musse, S.R., Thalmann, D.: A Behavioral Model for Real-Time Simulation of Virtual Human Crowds. IEEE Transactions on Visualization and Computer Graphics 7(2), 152–164 (2001)
91. Emering, L., Boulic, R., Molet, T., Thalmann, D.: Versatile Tuning of Humanoid Agent Activity, Computer Graphics Forum
92. Bouvier, E., Guilloteau, P.: Crowd Simulation in Immersive Space Management. In: Proc. Eurographics Workshop on Virtual Environments and Scientific Visualization 1996, pp. 104–110. Springer, Heidelberg (1996)
93. Bouvier, E., Cohen, E., Najman, L.: From crowd simulation to airbag deployment: particle systems, a new paradigm of simulation. Journal of Electrical Imaging 6(1), 94–107 (1997)
94. Brogan, D., Hodgins, J.: Robot Herds: Group Behaviors for Systems with Significant Dynamics. In: Proc. Artificial Life IV, pp. 319–324 (1994)
95. Ulicny, B., Thalmann, D.: Crowd simulation for interactive virtual environments and VR training systems. In: Proc. Eurographics Workshop on Animation and Simulation 2001. Springer, Heidelberg (2001)
96. Cremer, J., Kearney, J., Papelis, Y.: HCSM: Framework for Behavior and Scenario Control in Virtual Environments. ACM Transactions on Modeling and Computer Simulation 5(3), 242–267 (1995)
97. Boulic, R., Becheiraz, P., Emering, L., Thalmann, D.: Integration of Motion Control Techniques for Virtual Human and Avatar Real-Time Animation. In: Proc. ACM VRST 1997, pp. 111–118. ACM Press, New York (1997)
98. Pettre, J., de Heras, P., Maim, J., Yersin, B., Laumond, J.P., Thalmann, D.: Real-Time Navigating Crowds: Scalable Simulation and Rendering. Computer Animation and Virtual Worlds 16(3-4), 445–456 (2006)
99. Treuille, A., Cooper, S., Popovic, Z.: Continuum crowds. In: Proc. SIGGRAPH 2006, pp. 1160–1168 (2006)
100. Morini, F., Maïm, J., Yersin, B., Thalmann, D.: Real-Time Scalable Motion Planning for Crowds. In: Proc. CyberWorlds 2007 (2007)

2

Intelligent Virtual Humans with Autonomy and Personality: State-of-the-Art

Zerrin Kasap and Nadia Magnenat-Thalmann

MIRALab, University of Geneva,
Geneva, Switzerland
{zerrin.kasap,thalmann}@miralab.unige.ch

Abstract. Intelligent virtual characters has been subject to exponential growth in the last decades and they are utilized in many applications areas such as education, training, human-computer interfaces and entertainment. In this paper, we present a state-of-the-art of virtual human mentioning about the use of intelligent decision technologies in order to build virtual human architectures. We consider various aspects such as autonomy, interaction, personification and presence. Each of these aspects comes to prominence in different applications although there is no slight distinction. This survey provides a novel insight to the current state of designing and modeling virtual humans using different decision technologies and can be used as a basis for several future directions.

2.1 Introduction

The metaphor of intelligent and human-like computer characters has been around for a long time and they are the result of the convergence of several fields such as computer graphics, computer animation, artificial intelligence, human-computer interaction and cognitive science. It also has close relationships to robotics since it can share the same know-how in order to model the cognitive behaviour of autonomous individuals. The impetus of the area also comes from the variety of application areas from training/education systems to human-computer interfaces and entertainment films/computer games. Each of these application areas requires different properties at different levels such as autonomous behaviour, natural language communication, recognition of real people, personality modeling, emotional behaviour, adaptation to environmental constraints, user needs, intentions and emotions. In order to build such virtual human architectures, several intelligent decision technologies are utilized such as artificial neural networks and hidden-markov models.

The paper is organized as follows. First, we mention about autonomous behaviour of virtual characters considering both the internal state of the virtual human and the state of the virtual environment. Several issues such as perception, decision-making, adaptation, action selection and planning are explained. Second section represents the importance of interaction capabilities of virtual humans by giving examples of current Embodied Conversational Agents (ECAs)

N. Magnenat-Thalmann et al. (Eds.): New Adv. Virt. Hum.: Arti. Intel., SCI 140, pp. 43–84, 2008.
springerlink.com

and mentioning about different components of interaction such as facial expressions, gestures and dialogue. In the third section, another important factor of virtual characters called personification is considered, first referencing to some psychological models of emotion and personality and later to some examples of computational models considering the effects of emotion and personality on the behaviour and apperance of virtual human. In the next section, we mention about the importance of feeling of presence considering the interaction between real and virtual human to make the experience more believable. Finally, we conclude with a summary of our discussion in this paper.

2.2 Autonomy

With the occurence of real-time applications such as computer games and human-computer interfaces, autonomous behavior of virtual characters has become to be important. In [172], Tomlinson mentions that traditionally animation was produced by specifying every frame in the mind of the creator and the characters do not have the ability of self-controlling their actions. This kind of technique is applied in many famous films like Shrek and Happy Feet. However, in a real-time application characters should decide what to do at a specific time by themselves based on some pre-defined rules like in the well-known game The Sims.

Before going into the literature of autonomous virtual humans, we first want to give the definition of the terms autonomy and behavior. Autonomy is defined as the self-governing of one's actions and acting independent of someone's control. Two reasons to have autonomous behavior is defined in [190]: fidelity of the virtual life to the real one and less work load for the designer since it is a long and difficult process to design every frame exactly. Behavior refers to the actions or reactions of an object or organism, usually in relation to the environment [194]. Earliest examples of behavioral animation can be found in Reynolds [145] and Tu [177] where they modeled the behavior and perception of animals such as birds and fishes. The complexity of the behavior of an organism is related to the complexity of its nervous system and organisms with complex nervous systems have a greater capacity to learn new responses and thus adjust their behavior [194]. For this reason, behavioral animation of virtual humans is a complex topic requiring analysis of cognitive and emotional processes. We will mention about three important levels in autonomy of virtual humans in the following sections: perception, decision-making and action.

2.2.1 Perception

In real world, people as autonomous agents only perceive a part of the environment where the conditions are unpredictable and changing continuously triggering a behavior in the human being. In [20] it is mentioned that doing the same thing for virtual environments requires having the same limitations as real in order to convince the user. In other words, virtual environments are created by a designer and all information about the virtual environment is available for an agent e.g. through a scene graph. Human like perception of the environment needs to be simulated to apply these constraints.

Earlier examples that models visual perception of autonomous characters can be found in Reynolds [145] where bird boids were able to perceive the distance to other boids. Synthetic vision is first introduced by Renault [112] where an actor navigates in a corridor avoiding obstacles by off-screen rendering of the scene from the point of view of the agent. This technique later applied by Tu and Terzopoulos [178] in the perception system of artificial fishes. The technique is based on ray-casting which means simply sending and checking the rays are reaching to which objects. Blumberg [169] used synthetic vision for his autonomous virtual dog in order to avoid obstacles. Noser et al. [168] proposed a synthetic vision model based on false-coloring and dynamic octrees. This model renders the character's point of view scene and the objects in the resulting image are assigned a unique color to distinguish between them. Octree structure is used as a visual memory representation calculating each pixel's 3D position from the 2D off-screen rendered image and from the depth information stored in z-buffer. Similar synthetic vision method is applied in [94] using a visual memory based on object IDs and their most-recently observed states. In [144] active perception of virtual soldiers is modeled based on binocular perspective projection of the color 3D world onto the animat's 2D retinas. In [137], synthetic vision model is extended with two different vision modes. In distinct mode false-coloring is applied to each distant object and in some cases objects are grouped according to some different criteria that is called grouped vision mode. A different approach to synthetic vision is proposed in [32] based on perceptual filters which receives a perceptible entity from the scene and decide if to pass or not. A main distinction between virtual human perception models is whether they are based on task perception or biological perception [75]. Biological perception or bottom-up visual attention is an alerting mechanism to unexpected dangers as explained in [138]. Task-based or top-down approaches such as [44] and [68] consider task-related objects as the focus of attention. In addition to visual perception, aural perception is also modeled and applied by some researchers. However there are few researches in this area when compared to visual perception. Lehnert [102] mentions about the fundamentals of auditory virtual environments. A recent example in [73] describes a framework to capture the position of the sound source and mention about some concepts related with human auditory perception such as interaural differences, auditory acuity and auditory filter.

2.2.2 Decision-Making and Adaptation

Decision-making and adaptation is the evaluation of what is perceived through reasoning, decision-making and choosing the appropriate action to perform which is often named as action-selection process. This evaluation is not only based on the physical properties of what is perceived from the environment but also it depends on various properties of the perceiver such as intelligence, past experiences, current motivations, plans, personality and emotional state. According to this interpretation of the environment, we decide what to do with the current situation and select among different alternative actions that leads us to a better state. This better state depends completely on the person and his/her priorities. Most of the studies separate the modeling of mind and actions from each other since human mind does not think of every detail of body motions such as "Now I need to make a step with my left foot" but rather focus on higher level goals such as "I will go to the supermarket". Improv system of Perlin [136] uses a three layer architecture where the lowest layer geometry of the characters are manipulated in real-time by an animation engine and high-level capabilities are controlled by behavior engine which simulates the mind. In SimHuman [189], there are two different modules such as motion controller and behavior controller which are part of the sense-decide-act sequence. In [3], low-level movement control such as facial expressions, motor control such as walking, running and grasping are handled by a separate module letting the agents share a common area for low-level actions. At the high-level each agent makes its own decisions in its mind and sends messages to low-level control process. Another example of synthetic characters that have a layered brain architecture can be found in [50].

Action selection problem occurs in dynamic environments consisting of intelligent agents and animats. Properties of intelligent agents are defined as reactivity, pro-activeness and social ability in [198]. They should be able to perceive their environment and respond in a timely fashion to changes that occur in the environment which is called reactivity. Pro-activeness is related with goal-directed behavior by taking initiative in order to satisfy the goals. They also show social ability by interacting with other agents. There are various approaches for decision-making (action selection) in intelligent agents ([198]). Earliest approaches to action selection are based on symbolic (deductive) reasoning and solved with logical deduction and theorem proving which is not suitable for real-time systems since it takes too much time and does not respond well to rapid changes. A well-known cognitive architecture based on symbolic reasoning is the Soar project [143]. The training expert Steve [146] uses the Soar architecture in its cognition module. Problems in the symbolic approach caused a new approach called reactive approach to appear which is more suitable for dynamic environments. The best-known reactive architecture is the sub-sumption architecture of Brooks [38] where behaviors are represented as simple if-then like structures and take place in a hierarchical structure where each agent has precedence over another [198]. Finite State Machines (FSMs) are also used as reactive architectures where affect of a condition on the current state results in a new state. Decision-making with FSMs was recently used in computer games such as Halo 2 [80] and Quake III [181] as mentioned in [1]. Probabilistic and fuzzy approaches are also used for reactive agents [82]. Another group of researchers combine deliberative and reactive approaches resulting in a hybrid architecture that consists of both symbolic models and reactive models [62] [121]. Goal-driven architectures are most widely used in virtual applications such as Belief-Desire-Intention (BDI) architecture. Belief stands for the agent's knowledge about the world, desires are the objectives to be accomplished and intentions are what the agents have chosen to do ([187]). BDI is based on practical reasoning [36] which is the process of deciding what to do in each moment of time in order to achieve the goals. An overview of BDI-style agents in the literature such as PRS, dMars and JACK can be found in [187].

Decision-making in virtual environments is applied in many other different applications. A constraint satisfaction framework for the planning of anytime agents in computer games is proposed and applied in the EXCALIBUR project ([122]). A more recent example is [43] where a decision-planning framework is developed for open-ended scenario environments where virtual characters face unexpected situations either from user interaction or reactions in the environment. The framework is based on ontologies representing the environment as interconnected concepts and it becomes easier to infer relations between objects. Most of the efforts on autonomous and intelligent agents focus on using machine learning algorithms for modeling the minds. A different approach is presented in [60], which creates the illusion of intelligence with navigation of dummy actors in intelligent environments. Intelligent environment idea is based on the ecological theory of perception which sees perception as what environment offers us - also called as affordances- rather than simple physical properties of the environment. In other words, the environment becomes intelligent through affordances of objects, places and actions similar to the smart object approach [86].

Living in a dynamic environment requires the ability of learning what to do in unexpected situations, which is also called adaptation. Perceptions can also be considered as acceptable or unacceptable according to the social rules and virtual one can learn not to show unacceptable behavior. Intelligence and emotions helps us in coping with unpredictable situations. Intelligence is rather related with reasoning, planning, problem-solving and requires abstract thinking ability to give logical decisions. Adaptation can also be achieved through governing emotions to stressful conditions. Personality is another important factor in learning. For example, a new experience can bother a neurotic personality but will be enjoyable for an extrovert and open person since he/she will be eager to learn new things. Agent frameworks without learning are composed of predefined perception-action rules, however adaptation requires to construct the behaviour rules in time learning through experiences. Learning approaches are connectionist approaches such as artificial neural networks and evolutionary approaches such as genetic algorithms. Learning through interaction with human user is rather a new area that is paid attention in the last years. For example a work done at MIT Synthetic Characters Group aims to train a virtual dog through clicker training in interaction with human using reinforcement learning technique [107]. AlphaWolf project from the same group focuses on social learning through emotions ([171]). Another example of emotional learning is the ALEC agent architecture [65] which has both emotive and cognitive learning and decision making capabilities using mainly neural networks. In [166] an autonomous virtual character learns to do a specific task such as jumping over obstacles using evolutionary algorithms. Imitation can also be used in learning, e.g. in [85] robots learn sensory-motor behaviors online by observing a person or a robot. In [54] emotional learning and imitation are combined enabling a virtual character to quickly adapt online due to interaction with human user by the way of learning through emotional feedback to maximize happiness. Learning is also important in communication between virtual characters in terms of more realistic interaction.

2.2.3 Actions

Movements of the human body can be investigated under two groups. One group of movements occurs during conversation such as facial expressions, gestures and posture changes. We will mention about these in more detail in section 3 which is about interaction with virtual characters. Realistic synchronization of these movements leads to more fidelity. Other types of movements are performed while doing an action such as bringing a glass of water or going to the post office. Realizing an action in a virtual environment requires the study of complex movements of the body such as walking, running and grasping an object. Believable interaction with the objects is necessary to realize most of these actions such as grasping an object in a realistic way or navigating without collision with rigid objects. Various techniques such as forward/inverse kinematics, physics-based animation and collision detection are used for more realistic animation.

Realistic synchronization of these actions is important as well as choosing the right action since sometimes it is possible to perform some actions concurrently and it is necessary to control the transitions between actions. For example, a person can be running while carrying something in his/her hand at the same time but can not be running and sitting at the same time. This parallel execution of actions is implemented in some studies such as PaT-Nets (Parallel Transition Networks) and HPTS (Hierarchical Parallel Transition Systems). PaT-Nets [125] are finite state automata running in parallel. The nodes of the net are processes and the arcs contains predicates and conditions which provides a non-linear animation model since movements can be triggered, modified and stopped by transitions to other nodes. HPTS architecture [97] is organized as hierarchical state machines giving priorities to certain behaviors.

In [67] it is mentioned that autonomous behavior should be integrated with user control in some systems. In order to believe that virtual humans are intelligent, we want them to understand our high level commands such as walk, open the door, go to school. In other words, given a goal by the human user, they should be autonomously realizing and reacting to it. Such kind of architecture is proposed by Badler et al. [156] in order to fill the gap between natural language instructions and activities. Parametric Action Representation (PAR) gives a desription of an action with properties such as performer of the action, objects used during the action or path followed, location, manner and purpose of the action [13]. A natural instruction such as "Walk to the door and turn the handle slowly" is represented as a parameterized form to be converted to animations with specified properties [204]. There are also other representation and scripting languages for character animation such as VHML, APML and RRL. A comparison of these languages can be found in [15].

A last notice about actions is related with the effect of emotions and personality on the actions. Emotional state can change the quality of motions when performing an action such as walking in an exhausted manner when unhappy. In addition, personality can bring different alternatives to the way of realizing a goal or motivation. For example an attentive person would prefer to prepare a nicely organized meal when getting hungry and a more practical personality can choose to eat a sandwich.

2.3 Interaction

For being convinced that a virtual character has a particular amount of intelligence, we would like to interact with them just like we do with real human. In order to realize this, an assembly of several components such as conversational abilities, facial expressions, hand-arm gestures and eye-gaze are required. Agents with these capabilities are called Embodied Conversational Agents (ECAs) [142]. Before mentioning about the techniques for the creation of human-like computer characters, we would like to mention about some examples of ECAs available in the literature and and their abilities so that it is possible to understand what is available currently and what will be the future directions.

One of the first examples of ECAs is REA [186], 3D embodied real-estate agent, which is capable of both multimodal input understanding and output generation through gesture recognition, speech recognition, discourse, speech planning and speech synthesis. REA uses a mixed initiative dialogue management pursuing the goal of describing the features of the house that fits the user's requirements while also responding to user's verbal and non-verbal input that may lead in new directions. She has the ability to track turn of speaking. When the user gives a sign of turn taking such as talking or doing a gesture, it gives the turn to the user and allow the user to interrupt her. She can initiate conversational repair when she does not understand what the user says and asks the user to repeat that part of conversation. When she realize that user is coming closer to her, she turns her face to the user and/or do some posture changes.

Another ECA that is developed by the same research group is BEAT (The Behavior Expression Animation Toolkit) [185] which is a first example of animation tools that uses natural language understanding techniques to extract the linguistic and contextual information contained in the text to control the movements of different modalities such as hands, arms, face and the intonation of voice. It realizes written text into embodied expressive behaviors just as text-to-speech systems realize written text into spoken language [185]. BEAT is capable of producing beat and iconic gestures. Beat gestures are formless hand waves that are observed when there is no additional information to produce more specific kind of gesture and iconic gestures are used to indicate surprise or unusual information [117]. Some of the other properties of BEAT are gesture suggestion

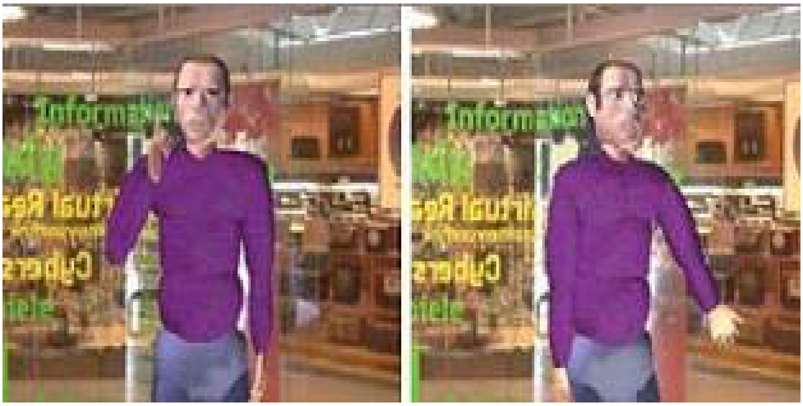


Fig. 2.1. MAX: Multimodal Assembly Expert

for contrasting concepts, raising of eyebrows as a signal of new material, gaze generation for turn-taking and intonation.

STEVE (Soar Training Expert for Virtual Environments) [146] is an animated pedagogical agent that teaches students how to perform procedural tasks such as operating and repairing equipment. Steve is capable of answering students' questions, showing each step of the tasks in interaction with the objects in the virtual environment. He uses several types of feedback such as gaze and pointing, to direct a student's attention or a nod of approval to show agreement with a student's action. Its capabilities are later on used in the Mission Rehearsal Exercise (MRE) [173] project which aims to prepare officers to face stressful conditions in foreign countries. In this kind of virtual environment there is more need for realistic characters with emotions and distinct personalities. The properties of Steve are extended so that he can produce speech depending on the personality and emotional state as well as the selected content. An expressive speech synthesizer is capable of speaking in different voice modes depending on whether the character is shouting, giving a normal speech or a command.

MAX (Multimodal Assembly Expert) [92] is capable of demonstrating assembly procedures to the user in an immersive virtual environment. He is capable of employing iconic and deictic gestures, eye gaze, emotional facial expressions and speech with intonation. All verbal and nonverbal utterances are created on-the-fly from XML specifications instead of utilizing predefined gestures from fixed libraries (Figure 2.1).

MACK (Media Lab Autonomous Conversational Kiosk) [182] is an ECA Kiosk that is able to answer questions about MIT Media Lab and give directions to its various research groups, projects and people. The goal of MACK is real-time multimodal input understanding/output generation and has the ability to reference a shared physical space with the user immersing into the physical world in order to give directions to the user.

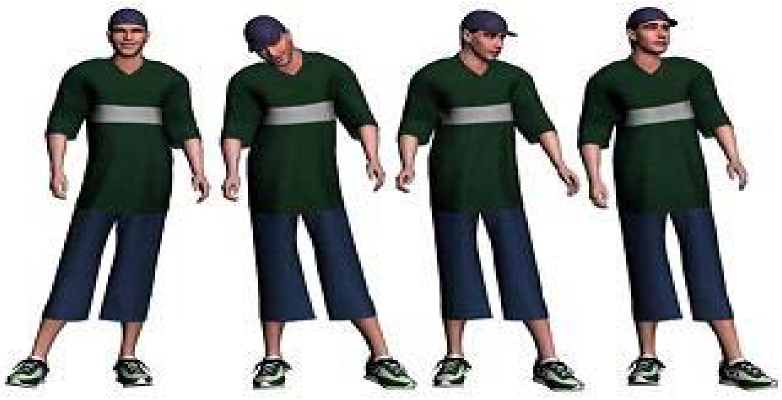


Fig. 2.2. Embodied Conversational Agent at MIRALab

GRETA [21] is an embodied conversational agent that is capable of multimodal interaction with the user through voice, facial expressions, gaze, gesture and body movements. Gestures are specified according to several properties such as type (iconic, beat etc), duration, assignment of single or both arms, timing of the keyframes relative to the duration of entire gesture, gesture phase (stroke, hold, retractain), arm position, wrist orientation and handshape with the help of a gesture editor. A gesture planner is used to find a mapping between the semantical structure of a particular piece of utterance and gestures in the database.

Another ECA developed at MIRALab (Figure 2.2) is capable of natural animation in real time with the consideration of idle motions such as posture changes from one resting position to another and continuous small posture variations caused by breathing, maintaining equilibrium [5]. This property is not considered in many ECAs so the result is not satisfactory in terms of realism and believability. In addition to idle motions and realistic body animation, MIRALab-ECA is capable of dialogue generation according to different emotional states that can be applied to different scenarios considering different personality models. Emotion and personality system effects the expressivity of both face and body in terms of different facial expressions such as happiness, anger, sadness and emotional idle motions. Although they are not directly related with interaction, some other properties such as real time cloth and hair simulation increase the believability of interaction.

In this section, we classify the techniques for the development of ECAs into three categories: facial expressions, gestures and dialogue. Facial abilities of ECAs can be classified as speech, emotional expressions and gaze. Speech and gaze (as a function of discourse) will be explained in this section. Emotion and personality modeling will be examined in the next section since it is a broader topic in itself with all its affects on face, body, gaze and voice. At last, we will mention about dialogue management techniques and applcations.

2.3.1 Facial Expressions

Computer animation techniques are evolving more and more in order to create more realistic virtual characters. Face is the first point of attention when one looks at an embodied agent and this is the reason that there have been lots of effort to improve facial animation techniques. [42] defines facial animation techniques under three categories. A similar categorization is also done in [88]. First method is based on manually generating key-frames and interpolating between them. This is done by manipulating the facial mesh at a low level in order to create morph targets or keyframes and interpolating between them. This technique is especially used in film industry since it gives very good results as a result of artistic work applied on a particular model. Second method of facial animation is based on automatically extracting visemes - visual counterparts of phonemes - from written text with the help of a text-to-speech tool or from acoustic speech with the help of speech recognizer techniques in order do synthesize speech with lip-synchronization. Third technique of facial animation is based on parameterization of the facial mesh with certain feature points, representing the captured data according to these feature points in order to analyse and synthesize. Second and third techniques of facial animation will be explained in more detail in the next pharagraphs.

Earlier work on computer generated animation of faces starts in 1972 by Parke [134]. A real model is painted a mesh on his/her face and animation is created by photographing freezed expressions. Later, Pearce et. al. [10] developed a system that enables the manual entrance of a phonetic script that would result in synchronized lip movements. Automated methods of lip-synchronization started in the second half of 1980s and followed two paths: computer synthesized speech and recorded speech. The former involves automatic generation of speech animation from typed text. Earlier work done in this field can be found in Hill et. al. [49]. Various facial animation systems that use different text-to-speech systems are developed later on such as [191] and [69]. With the advances in text-to-speech (TTS) systems both produced commercially (e.g. Microsoft, IBM, ATT) and by research groups (e.g.Festival [29]) lead to easier linguistic representation of a given text in forms of building blocks such as phonemes, words etc. Phoneme timing information coming from a TTS system is used for mapping phonemes to visemes which are predefined and stored in a viseme database. Steps in creating speech animation are well defined in [88]. Building blocks of speech animation coming from a viseme database are interpolated according to the timing information in the phoneme stream coming from the TTS system. In order to obtain realism in speech animation, simple interpolation between phonemes is not enough since each phonetic segment is influenced by its neighboring segments which is called coarticulation. Approaches for computing the effects of coarticulation can be found in [135] and [46]. Although computer synthesized speech from text is good for providing accurate synchronization between speech and lip movements, it is still lack of properties such as natural rhytm, articulation and intonation provided in natural speech [104].

For more realism it is also possible to extract phoneme information from recorded speech although it has its own disadvantages. Specifing phoneme timings from natural speech is a challenging task and requires the use of different techniques for extracting the parameters in a speech signal such as Linear Predictive Coefficients, Fourier Transform Coefficients, Mel-Cepstral Coefficients as well as pitch and energy [88]. Machine Learning techniques such as Hidden Markov Models (HMMs) and Neural Networks (NNs) are used to train the processed audio and processed visual data or statistical techniques such as Principal Component Analysis (PCA) are used to analyze the parameters in recorded speech ([120], [56]). In [34], Voice Puppetry system is presented for the generation of face animation from expressive information in an audio track. [200] describes a video realistic speech animation technique using a small video corpus.

Third method of facial animation is based on the parameterization of the facial animation. Previously, facial animation was being done by manipulating the facial mesh at a low level and interpolating between keyframes. However, it has some disadvantages since the development process is slow and needs artistic efforts of designers. Once an animation is created for a particular face it can not be applied to other facial meshes. Techniques that can automatically and easily produce facial animation and that can be applied on any facial model are important for an ideal facial animation system. Parameterized models address this problem and they allow to generate facial expressions by manipulating a set of parameters. Animations are specified in terms of these parameters so that an animation can be applied on another facial mesh that contains the same feature point information. Earliest parameterization technique for facial animation is the Facial Action Coding System (FACS) developed by Ekman and Friesen [64]. The goal of this work was to make a classification of facial expressions based on the muscles that produce them and the standard is found very useful by psychologists and animators and became popular. FACS contains 44 action units (AUs) and facial expressions are generated as a combination of different AUs. Another study developed by Kalra [130] used the Minimal Perceptible Action (MPA) similar to FACs but with some extensions such as asymmetric facial movements and head nod. Recently, MPEG-4 facial animation standard is accepted and widely used for 3D facial animation design. Facial Definition Parameters (FDPs) are feature points that are used to characterize the face, in other words they define what a face is. Facial Animation Parameters (FAPs) are used to define an animation to produce faces with speech, expression and emotions. Each FAP value corresponds to the displacement of one feature point on the face in one direction in terms of FAPUs (Facial Animation Parameter Units). FAPUs are calculated according to the given model and they are the fractions of key facial distances such as the distance between two lip corners or between eyes. More information about MPEG-4 facial animation standard and its implementations can be found in [16]. Feature point based geometric deformation methods are used for the animation of parameterized facial meshes as described in [149]. First, a mesh with control points is defined with the information for each feature point indicating in which direction the movement of the feature point is constrained. Second, given a geometric mesh with control point locations, regions of influence by each of the control point are calculated. Each vertex of the mesh should be controlled by not only the nearest feature point, but other feature points in the vicinity, in order to avoid unrealistic animation. The last step involves calculating the actual displacement of all vertices of the mesh in real-time from the displacements of feature points. Other examples of parameterized facial mesh deformation can be found in [98], [9] and [150]. There are also various other methods of facial animation such as finite element method, muscle based modeling, pseudo muscles, spline models and free-form deformations. A survey of facial deformation techniques can be found in [124].

In order to build realistic facial model, eye movement is also important as well as face geometry, simulation of facial muscles and lip synchronization. The importance of eye engagement during social interaction and discourse is mentioned in some studies such as [27], [78] and [74]. Gaze can be used as a signal of paying attention [17]. While listening, turning the gaze to the speaker allows one to see the speaker better and during speaking, speaker wants to be better seen by the listener and turns his head to the listener. This means that gaze is not only intentionally used to collect information but also interpreted as a signal of attention to what is spoken. Gaze also has a regulatory function during conversation. For example, it takes role in conversation initiation since there is a need for engagement between speaker and listener at the beginning of the dialogue. At certain points of talk, listener's gaze can cause the speaker to change the direction of the talk and take a role in turn-taking by signaling a start or end of a turn. Recently there are some studies that developes gaze models in order to convey emotions and impression and some research is focused on integrating gaze models in immersive virtual environments where multiple agents interact with each other in order to increase the feeling of presence. We will mention about these studies in the next sections about emotion and presence.

2.3.2 Gestures

Gesture is a form of non-verbal communication that causes a particular change in the shape of the arms and body in order to convey complementary information to discourse. Gestures are different from other movements of body such as grasping and reaching in that they carry the semantic characteristics related with the content of the speech [91]. In order to develop computational models of gestures, some form of parameterization is required for the description of qualitative aspects of gestures. McNeill's approach [117] to gestures is based on psychology and emprical studies and he classifies gestures into several categories such as iconics, metaphorics, deictics, beats and emblems. Computational methods on gestures such as BEAT [185], REA [186], GRETA and MAX are mostly based on the gesture model of McNeill. A different approach to parameterization of gestures is applied in the EMOTE (Expressive MOTion Engine) system [203] which usues effort and shape parameters of Laban Movement Analysis (LMA) [96] in order to generate more natural synthetic gestures. Classification of gestures is done in the domain of movement observation independent of their communicative meanings since there can be movements that can not be classified as gestures or they can be perceived as a particular gesture in one culture but not in another culture [203].

The main problem with realistic gesture generation is the synchronization of the gestures with speech. Studies in believable gesture generation can be grouped into two directions. One group is concerned with semantic aspects of gestures such as timing of gestures. Automatic prediction of gesture timing from synthesized speech has been studied in some systems such as BEAT [185], REA [186] and MACK [182]. In [117], four different phases of gesture is defined: preparation, stroke, hold and retraction. Stroke is the mandatory part of a gesture that carry the meaning. Usually the solution with timing of gesture is matching the stroke part of the gesture with the most emphasized part of the utterance. [90] describes a study that extends the work in BEAT , REA and MACK where the form of gesture is derived on-the-fly without relying on a lexicon of gesture shapes or "gestionary". Another more recent example is GESTYLE [147] that brings the notion of style in forms of clothing (formal/informal), choice of language (polite/casual), gesturing motion characteristics (expansive/subdued), gesturing frequency, gesturing repertoire, characteristics of speech (intonation, volume) and so on. Style is defined in terms of when and how the ECA uses certain gestures and how it modulates speech.

Second group of researchers are more concerned with realistic generation of gesture movements. The quality of the resulting animation in the above mentioned systems is limited since the animations are generated procedurally only for a few joints resulting a mechanic animation. In nature, a gesture is not just composed of the movements of a few joints but influences of each movement of a joint on other joints should also be considered. In [5] an animation synthesizer that allows the generation of small posture variations and posture shifts is described. Generated idle motions are based on statistical data obtained from a large set of recorded animations using the Principal Components Analysis (PCA) method. Another research in [101] describe a method for using a database of recorded speech and captured motion to create an animated conversational character. Motion samples are combined with new speech samples and they are blended phrase-by-phrase into extended utterances. Other examples of motion synthesis can be found in [126], [95] and [199] where they use large databases of motion segmented and blended together. In [127], a motion synthesis approach is presented that allows the consideration of annotation constraints so that the system assembles frames from a motion database according to the timing information for each action. [165] presents a real time animation system that extracts the rhythmic patterns of motions such as dance and marching and synthesizes a novel motion in an on-line manner while traversing the motion transition graph, which is synchronized with the input sound signal and also satisfies kinematic constraints given explicitly and implicitly. In Style Machines [35], the goal is stylistic motion synthesis by learning motion patterns from a highly varied set of motion capture sequences.

2.3.3 Dialogue Management

Natural language dialogue is a very important part of the interaction between human and machine. Dialogue management systems basically can model a dialogue between a human and a computer but there are also systems that can realize the conversation between two or more computer characters and human beings. A believable dialogue system requires the integration of the functionalities such as response generation according to emotional state, allowing interruptions, repairing of dialogue, feedback and turn-taking.

Some features of discourse and difficulties of modeling dialogue are mentioned in the survey of dialogue management systems [19]. A dialogue has an opening, body and closing and although the user takes control in most parts of the dialogue, the overall dialogue should be the result of a mixed initiative between user and agent. In dialogue it is common that we use incomplete sentences in order to explain something in short or as a result of our speaking style. This incomplete information should be recovered from the context of the dialogue. A recovery mechanism is also required when one side of the conversation can not understand the other one and it is the case with computers since they are not able to understand every word we say. Another feature of dialogue is about indirectness where cognitive skills are required to understand the overall meaning of dialogue. Turn-taking requires deciding when one of the speakers start talking or give turn to the other speaker and it becomes more complicated when there are more than two speakers. The use of fillers such as the words 'a-ha' and 'yes' are important in order to give feedback to the other speaker to show that you are paying attention to what he/she is saying. Non-spoken period of speech should also be interpreted as a part of the dialogue for more realistic conversation.

In [175] four approaches to dialogue systems is defined: (1) finite-state based and frame-based systems (2) information state and probabilistic approaches (3) plan-based approaches and (4) collaborative agent-based approaches. In finite-state methods dialogue is composed of a sequence of predetermined states and flow of dialogue is determined by transitions between states. Frame-based approach is an extension of finite-state based approach addressing the problem of flexibility. The user is asked questions in order to fill in the slots for a given template related with a task such as a train timetable [118]. Information state approach is an effort to overcome the limitations of finite-state based and frame-based approaches. It is composed of informational components such as participants, linguistic and intentional structures, beliefs and their formal representations such as lists, sets, records and so on. An information state is updated through dialogue moves based on update rules and update strategy. Plan-based approach is more complex than the previous approaches and originates from the idea that humans communicate to achieve a particular goal. Collaborative approaches or agent-based dialogue management approaches are based on viewing dialogues as a collaborative process between intelligent agents. Both agents work together to achieve a mutual understanding of the dialogue.

ELIZA [192] is one of the first attempts for dialogue generation which is based on pattern-matching techniques that allows generating standard answers to certain word combinations in a given phrase. Another well known program, ALICE [12] which is an extension of ELIZA utilizes an XML based language AIML - Artificial Intelligence Modeling Language. In [57] a dialogue model similar to ALICE is presented but with some extensions to consider different emotional states. A parallel Finite State Machine (FSM) algorithm is applied where several FSMs run concurrently and each FSM represents a dialogue unit about a certain topic.

One of the more recent examples of dialogue systems is TRINDIKIT framework [174] which focuses on the development of new technologies for adaptive multimodal and multilingual human-computer dialogue systems. TRINDIKIT is a toolkit for building and experimenting with dialogue move engines and information states and is an example of information state approache that is mentioned in the above paragraph. Information state means the information stored internally by an agent and a dialogue move engine updates the information state on the basis of observed dialogue moves and selects appropriate moves to be performed [55]. A similar study to TRINDIKIT is DIPPER framework [103] for building spoken dialogue systems, based on the information state theory of dialogue management and it is utilized in the GRETA embodied agent [33]. GRETA also utilizes the BDI model [197] of Belief- Desire-Intention which is used in order to simulate the mind of the ECA. BDI model is also integrated with other dialogue systems such as [37] and [30]. For a good classification of various applications and more detailed information about state-of-the-art of dialogue systems, we refer to [175], [118], [154] and [201].

2.4 Personification

Personification means giving human properties to non-human objects and this issue is becoming to be important in the world of virtual humans in the last years. While looking at the perspective of traditional intelligent systems such as expert systems or decision support system, having emotions can be seen as a non-desirable property. However, this is not the case for the domain of believable agents since we prefer them to behave as human as possible. Social behavior of computer characters with emotion and personality increases the realism and quality of interaction such as in games, story-telling systems, interactive dramas, training systems and therapy systems. They also can replace many of the service areas that real people are employed such as a museum guide or a receptionist.

In [63], a social robot that is placing a receptionist in a hotel is presented. The robot has a LCD head that displays a highly expressive graphical face and is capable of conveying different moods such as neutral, positive and negative. It is observed that people prefer to interact with negative receptionist for shorter periods. Emotions are also both a way for coping with stressful factors in the environment and this property is important in many computer systems with the purpose of training and therapy. Coping is defined as a phenomenon that determines how one responds to the appraised significance of events e.g. an undesirable but controllable event motivate people to reverse the bad conditions

[115]. For example in the Mission Rehearsal Project [114], lieutenants learn to cope with dramatic situations such as in a war and improve their decision-making capabilities under stressful conditions.

When we talk about personification, we usually consider two factors: Personality and Emotion. Personality is a phenomenon that makes it available to distinguish between different people. This is also the case in our interaction with virtual characters. When we are immersed into a virtual environment populated with virtual humans, experiencing that they all behave differently under same conditions, increases their believability. Emotion is another major component of personification since in real life emotions affect people's all cognitive processes, their perceptions, beliefs and the way they behave. The importance of emotions as a crucial part of virtual humans is previously mentioned in [139].

Empathy is another property of personified virtual characters and a key factor to increase believability. On one side a virtual character can behave in a way that leads the user to establish emphatic relation with it or the agent itself can be emphatic and behave in an emphatic way towards other agents and towards the user [8]. FearNot! [8] is a project addressing the bullying problems in schools letting the children users to establish emphatic relations with the characters in the virtual environment. There are many other issues in the personification of virtual characters. In [161] they ask the question if it is possible to develop friend-ECAs so that interaction with that character or learning from it becomes more fun. Humour is another important property of real people and can be very useful in human-computer interaction increasing the level of joy and this leads to more cooperation between user and agent [123].

Several models of emotion and personality are developed in the domain of psychology. However, these models are usually produced for the purpose of psychological studies rather than being used in the creation of computational characters. This gap between the models of emotion and personality and the complexity of modeling human feelings will keep this area as the focus of attention for many computer scientists for the future. In the next sections, we will first examine different models of personality and emotion in the literature. Second we will mention about some examples of virtual characters that are developed based on these models. The effects of emotions on different parts of a character's body (e.g face, body and gaze) will be considered in addition to their effects on the character's actions and decisions.

2.4.1 Personality

Personality influences the way people perceives their environment and affects their behaviors and actions and distinguish people from one another. Although there is no universally accepted theory of personality, Five Factor Model or OCEAN model [116] is most widely used one in the personality simulation of computer characters. According to this model, personality of a person can be defined according to five different traits: openness, conscientiousness, extroversion, agreeableness and neuroticism and they are explained in [71]. Openness means being open to experience new things, being imaginative, intelligent and creative whereas conscientiousness indicates responsibility, reliability and tidiness. In other words, conscientious people think about all the outputs of their behaviors before taking action and take the responsibility. An extravert person is outgoing, sociable, assertive and energetic to achieve his/her goals. Agreeableness means a person is trustable, kind and cooperative considering other people's goals and is ready to surrender his own goals. At last, a neurotic person is anxious, nervous, prone to depression and lack of emotional stability. Usually, a character is represented as a combination of these traits possibly with emphasize on one of them. Although this static trait-based model of personality does not reflect the complexity of human behavior truly, it has been widely accepted to be used in computational models because of its simplicity.

2.4.2 Emotion

There is a variety of psychological researches on emotion and they are classified into four different theoretical perspectives: Darwinian, Jamesian, cognitive and social constructivist approaches [47]. Darwinian perspective is based on the research of Charles Darwin in 19th century and defines emotions as functions we use for coping with difficulties and mentions that they evolved as a result of natural-selection. This perspective is focused on physical displays of emotion on face and body. Jamesian perspective of William James in 1800s is based on the idea that emotions always occur with bodily changes such as posture changes or facial expressions. Cognitive approach to emotion was formulated recently in 1960s and explains emotions as a result of cognitive appraisal of the environment. Social constructivists believe that culture is an important factor while experiencing emotions and occurs as a result of social rules. This is a more recent approach that appeared in 1980s. In order to create computational models of emotions we need some annotation mechanisms. Ekman [59] defines six basic labels for emotions: fear, disgust, anger, sadness, surprise and joy. He follows the Darwinian approach to emotions emphasizing the universality of human emotions.

Recently, cognitive appraisal models of emotion are becoming to be more preferred since they better explain the overall process of how emotions occur and affect our decision-making. Appraisal means a person's assessment of the environment including not only current conditions but past events as well as future prospects [79]. Examples of these models are of Lazarus [99], Roseman [79] and most popularly used OCC model of Ortony, Clore and Collins [7]. In OCC model, agent's concerns in an environment are divided into goals (desired states of the world), standards (ideas about how people should act) and preferences (likes and dislikes) [110] and twenty-two emotion labels are defined.

In addition to the four theoretical approaches to emotions, there are also some studies on emotions based on neurophysiology. The first representative of this approach is Plutchik [141] who defines eight emotions in an emotion spectrum like the color spectrum and it is possible to produce new emotion labels by the way of mixing these emotions (e.g. disappointment = sadness + surprise). Other models of emotion such as activation-evaluation [193] defines emotions according to abstract and continuous dimensions rather than discrete labels. In the activation-evaluation cycle, the vertical axes show the activation level (an emotion's activation level is in somewhere between very active and very passive) and the horizontal axis shows how the emotion is evaluated in terms of being positive or negative.

2.4.3 Applications to Virtual Human

Various models and applications are developed that uses the above mentioned psychological models of emotions and personality. [23] describes what the OCC model of emotions is able to do for an embodied emotional character and what it does not. Emotion processing is split into five phases: Classification, Quantification, Interaction, Mapping and Expression. In the classification phase the character evaluates an event, action or object and affected emotional categories are found. Intensities of affected emotional categories are calculated in the quantification phase. In the interaction phase the emotional value interacts with the current emotional categories of characters. In the next phase, twenty-two emotional categories are mapped to a lower number of categories and at the last step the emotional state is expressed through facial expression and can influence the behavior of the character. However this model is not enough to explain all properties of a character and needs to be extended considering other factors such as personality, emotional history and social rules. The importance of another factor, namely mood is realized recently in simulation of affective states of human being. Mood is a prolonged emotional state caused by the cumulative effect of momentary emotions [196] and is a dynamic property that changes with time. People are often defined to be in good or bad moods and it is possible to have cases such as smiling although being in a bad mood. In [93] a generic model is described for emotional conversational virtual humans with personality. In this work, an individual is represented as a combination of three factors at a specific time: personality, emotion and mood, where emotion and mood are dynamic properties and personality is a static property. Emotional state of a person is updated in two steps: mood update and emotion update. Mood is updated with factors such as personality, emotion history, mood history, emotion influence and a decay function. Emotion influence is calculated according to the dialogue state using the OCC appraisal model. Emotional state is updated according to the new mood history and considering other factors mentioned in the first step. A relation is constructed between the goals, standards and attitudes in the OCC model and OCEAN model of personality. For example, an agreeable person can adapt his goals to other people or abandon his goals in favor of others or an open person is prone to fast changes in standards. For the visual front-end the 22 emotions of OCC model is extended with two other emotions surprise and disgust that does not take place in the OCC model and grouped into six basic facial expressions of Ekman. In Figure 2.3 a virtual character from MIRALab with emotional expressions can be seen.

Personality, mood and emotions are used in many applications as a three layer model of personification where mood is considered as a medium-term property of personality between temporary emotions and permanent personality traits.

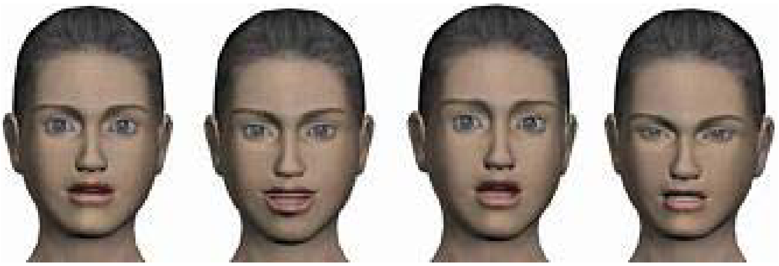


Fig. 2.3. Emotional facial expressions

ALMA (A Layered Model of Affect) [66] aims to provide a personality profile as well as real-time emotions/moods and uses Big Five personality model and OCC emotions. A character's personality profile is composed of personality definition and appraisal rules defining how a character appraises its environment. In [163] a platform is developed for dynamic character design for storytelling environments where storytelling players or designers have an effective way of controlling synthetic characters through high-level personality and emotion controlling. Abridged Big Five Circumplex Model (AB5C) is used as a personality model since the five factor model (FFM) is found insufficient and emotions are represented with Ekman's six basic emotion labels. In BASIC [4], a believable, adaptable socially intelligent agent is developed that extends the three layer personality model of emotion-mood-personality with memory and social cognitive factors. Personality system receives events as stimuli to internal state and processes them according to the current emotional state, mood, memory, personality and social cognitive factors. In this model characters that have a previous experience with other characters will be influenced by their previous interactions with the consideration of memory. Social cognitive factors are regulators that act across the three layers. For example they regulate the speed at which a characters emotions, mood and memory will return to neutral over time or specify how much past experiences affect the current mood.

An application of emotion representation using activation-evaluation space can be found in [11]. The reason for preferring this model is that it is possible to capture the emotional diversity in the discourse and conversion to discrete descriptions of emotions is also possible. Another group of researchers believe that psychological models in the literature are not good at studying the contextual factors of multimodal emotional behavior. In [109] a context specific multimodal corpora of TV recordings of real human emotions is used instead of motion captured data in order to study the coordination between modalities during complex emotional behaviors. The annotation and modeling of emotional behaviors require representing the multiple levels of abstraction and temporality in the emotional process.

In [148] an anthropometric agent, Max is developed whose cognitive functions are modulated with an emotion system and emotional state affect facial expressions, gesture and speech. The underlying emotional system consider emotion, mood and boredom factor that occurs in case of absence of stimuli. As a result of an annoying conversation with the user, Max gets angry with his face, gaze and body and leaves the screen when the user goes on annoying and turns back when he calms down. In [167] a knowledge-based system is described to store animations of reflex movements taking into account personality and emotional state.

Considering the visual front-end of computational systems that contains a personality model, earlier applications started with a 3D talking head capable of facial expressions and speech responding according to the dialogue state and personality model. At the next step, more has to be considered such as affects of emotions on body movements, gaze, voice and physiological signals such as blushing and sweating as a result of emotional process. In addition personality model, should take part in the autonomy of the character and take role in the action-selection process of the agent.

Effects on Body. Efforts in recognizing the mood of a person are mainly focused on facial and oral cues but body movements are not considered that much. The reason for that are the lack of systematic study on gesture features and the high variability of the possible gestures [131]. In [52] it is mentioned that the clues of emotion can sometimes be more easily understood from the body movements. They studied the photographs of fearful and angry faces and bodies to create face-body compound images either matched or mismatched and found that when face and body convey conflicting emotional information, judgment of facial expression is hampered. Most of the studies on emotional body movements are based on dance motions. However in [131], it is indicated that dance motions are exaggerated when compared to natural gestures and postures of body and they present their approach to address this problem. In the proposed technique, they apply a set of motion capture experiments on a group of people expressing a given emotion and another group of people label the movements. Each frame is defined according to 24 posture features such as the distance between right hand and right shoulder etc. This information is calculated through projection of 3D data onto 3 orthogonal planes. Gestures are grouped according to the emotion labels and PCA analysis is applied on each group in order to remove the outliers and to obtain principle components of each emotion. Egges [2] presents an animation system which allows the synthesis of emotional balance shiftings according to the emotional state using the activation-evaluation [193] model of emotions (Figure 2.4). For each posture an interval is specified according to the activation-evaluation circle and appropriate animation segments are chosen automatically by the animation system given an emotional state. In addition pause length between different postures is also adapted according to the activation level since higher activation level will result in shorter pauses and more shifts.

Effects on Gaze. The importance of eye movement is mentioned in the previous section as a function of social interaction and discourse. Gaze also has an expressive function and is used to convey information such as liking - continuous gaze indicates liking [108]. In [183] it is mentioned that personality is recognizable by

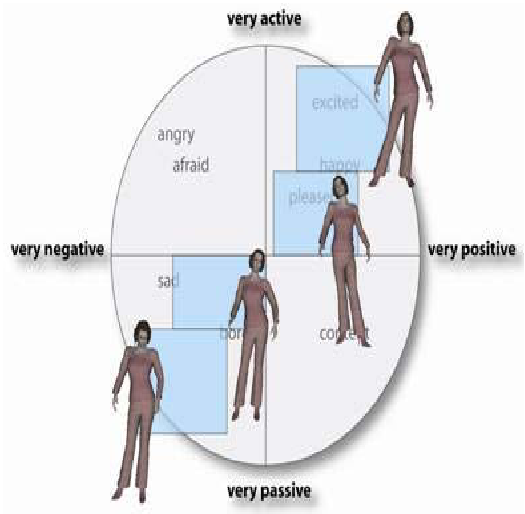


Fig. 2.4. Emotional idle motions mapped to activation-evaluation disk

some personality cues such as head node and eye gaze. They mention about two functions of eye gaze: monitoring functions (to collect information about other person) and expressive functions (transfer information like feelings and attitude). Although expressive functions give more cues about personality, monitoring functions also play a role, e.g. a person that is not using monitoring functions can be perceived as not paying attention. The length and direction of gaze is also important. For example, people do not want to have constant mutual gaze during conversation. More continuous gaze results in higher ratings for liking and activity. Quick eye movements indicate nervousness or anxiousness. In [108], a gaze movement model is developed that enables an embodied interface agent to convey different impressions to users. Three parameters from psychological studies is considered in the gaze model: amount of gaze, mean duration of gaze and gaze points while averted. They implement a two state Markov model to implement the gaze model where one state is the gazing state and the other is the averted state. This stochastic model of eye gaze reduces the problem of periodicity and machine-like impression. Impressions such as like/dislike, strong/potency, warm/cold are validated through the change of parameters in the model. Another eye movement model is presented in [22], based on empirical studies of saccades and statistical models of eye-tracking data. First, a sequence of eye tracking images is analyzed to extract spatio-temporal trajectory of the eye. A statistical model of saccades is constructed. The images are segmented and classified into two modes: talking mode and listening mode. The model reflects the dynamic characteristics of natural eye movement, saccade magnitude, direction, duration, velocity and inter-saccadic interval. This information is combined with other facial functions such blinking, lip movements, head rotation. The model is tested to evaluate the impressions such as friendliness, engagement, interest etc.

Effects on voice. One aspect of naturalness missing in synthetic speech is emotions and this area has been started to be studied in the last decade. [153] describes a two axes emotional speech model where one axes is related with how a given emotion is expressed in speech, in other words the vocal correlates of emotions and the other axes is related with the properties of the emotional state. Some techniques for emotional speech synthesis are also mentioned in this study such as formant synthesis, diphone concatenation and unit selection.

Effects on Decisions and Actions. Emotion and personality has proven effects on cognitive processes such as decision making, learning, memory and planning which are concepts studied in different disciplines such as psychology and neuroscience. For human-like intelligence the role of emotions in decision-making is very important. Our emotions take role as motivational factors for our decisions and effect our actions. When looking at the perspective of autonomous agents, an agent can choose whether or not to pay attention and/or react to a given environmental stimulus according to its goals[40]. Emotions can change priority of the agent's goal by the way of affecting its decision-making process. They also provide a way of adaptation in dynamic environments for protection from environmental influences (e.g. fear) or blocking these influences (e.g. anger) [41]. Emotional behavior of an autonomous agent can also be understood as a signal that lets others to anticipate its possible behavior [41].

In [110], a framework for integrating cognitive appraisal and personality theory in decision-making is explained that is based on OCC model of emotions and "Big Five" personality model. One extension of this framework to original OCC appraisal is a mechanism for choosing between alternative decisions and courses of actions based on emotional intensity and personality weights. An example scenario of three terrorists with different goals and personalities is considered in this study. In [128], a prototype prey/predator environment is presented for decision-making. Preys are the emotional agents and they walk around in the environment trying to survive looking for resources, fighting predators and teaming up to fight together against predators. Agents also have different personalities: Hero and Grumph. Hero has an extravert personality, it is self-confident and optimistic. Grump has an opposite personality, easily appointed and introvert. Hero has self-control over his emotions except the very basic emotions such as fear and distress and these emotions has affect on the action-selection for both agents. Grumph's action-selection is highly influenced by negative emotions such as shame, reproach and resentment and he is less likely to attack predators. Another study on emotional action-selection [53] aims to define the interactions between motivations, emotion and personality to understand how they affect decision-making of virtual humans. According to this framework, emotions influence motivations at the qualitative level such as length, perception, activation and interruption.

Physiological Signals of Emotions. Emotions do not only come with the changes in facial expressions, body movements or gaze but also can reveal themselves with other emotional outputs of body such as blushing and sweating. There is little study that try to create models for these kinds of properties (e.g. [87]). However, there is more efforts on the recognition of physiological signals of emotions that can guide to develop models relating physiological states of emotional states. In [106] Galvanic Skin Response (GSR) and Electromyography (EMG) is used to map different emotions to an arousal/valence space. GSR is an indicator for skin conductance and increases linearly with a person's level of overall arousal. EMG measures muscle activity and it is correlated with negatively valenced emotions. There are also other signals such as respiration, temperature, heart rate and pupil dilation. Another interesting research that uses physiological signals of the body in order to convey emotions is an Emotional Wardrobe [159] that is concerned with expressing the inside to the outside. In this study, physiological sensors attached to an inner cloth are sent to an affective coder for assessment of the mood of the person and the result is reflected on an outer cloth where color-changes are used to convey emotions.

2.5 Presence

Presence is a key factor in the interaction between human being and virtual human in order to create the illusion of reality. From the application point of view virtual humans are taking roles in pedagogical learning systems, maintenance training systems, therapy systems, cultural heritage applications and many others for reasons such as elimination of the risk factors, cost-effectiveness or for experiencing something that is impossible to see in our time or just for entertainment purposes. Using these systems effectively, at the same quality with their real counter-parts and experiencing the same conditions as we are in the real world can be realized by focusing on the presence factor.

There is a variety of definitions and classifications of presence. Usually it is defined as the subjective sense of "being there". Lombard et. al. [39] defines presence as "the perceptual illusion of non-mediation" which means we believe in the illusion that the sensory stimuli created by technological devices are real. Heeter [72] defines three types of presence: (1) personal presence refers how much a person feels that he or she is a part of the virtual environment; (2) social presence is the feeling of other being's existence in the environment (3) environmental presence means feeling the reactions coming from the environment. Slater [157] mentions the significant distinction between the terms immersion and presence which are often used interchangeably. Immersion is an objective notion which can be defined as the level of sensory stimuli coming from a technical device such as data gloves. Level of immersion is measurable and comparable between different devices through some technical properties of the devices. Presence is a subjective phenomenon and it is based on different subjects' experiences of presence and occurs in their minds. For a more detailed definition and classification of presence, we refer to [180].

We will look at feeling of presence in virtual environments from two perspectives: one is related with how people sense the virtual environment through sensory channels and second is related with how user is realized by the virtual character. First one is related with at what extent people perceive the virtual environment as close as the real one through their sensors of vision, hearing, touching and even smelling and tasting. As much as the participants are able to use their sensors in the virtual environment similar to the way in real, they will be more immersed into the virtual. Sheridan [155] defines three dimensions of presence where one of them is the amount of sensory channels. Another dimension of presence is at what extent these sensory channels can be controlled and the last dimension is the ability to modify the environment. This kind of presence is accomplished by using some devices such as head-mounted displays, haptic devices such as data-gloves and force-feedback devices. By this way we can, for example, see the virtual objects and virtual human as real, touch them, manipulate the environment by changing the place of virtual objects or experience a touching feedback from the virtual world. This kind of presence is often named as physical presence by many researchers. Second perspective of presence is social presence and requires feeling the existence of other intelligence in the environment [100]. This means the intelligent character should realize us and give human like responds to us. Social presence brings the notions of face-to-face communication and most importantly recognition of the real environment and real people, in other words immersion of virtual one to the real.

In the next sections, we will mention about these two perspectives of presence considering the state-of-the-art technologies and applications. Biocca [28] mentions that feeling of presence in virtual environments is like a day dream occurring in our minds. In the future people might achieve to create immersion into virtual environment without using any physical device and just experiencing it in their minds even do not knowing this is a virtual one, like in the films Matrix and Vanilla Sky.

2.5.1 Physical Presence

People can experience the virtual environment using a computer screen or they can be fully surrounded by the virtual place what we call immersion. This is accomplished through special hardware that creates sensory connection (e.g. visual, auditory and tactile) between our body and the virtual environment. Although they are very rarely considered, smelling and tasting in virtual environments are also necessary for full immersion.

Visual Sensory Channel. Visual sensory channel is the one that is most studied since it is the basic step of immersion. It is necessary for seeing and navigating in the virtual environment. Today's most widely used display devices for immersion are head-mounted devices and projection-based devices (CAVE-like systems). HMDs are mounted on the head of a person and create an image in front of the person's eyes. First HMD is created in 1968 [164] by Sutherland using CRTs (cathode ray tubes) as a display technology and are capable of tracking the user's head position and orientation in a mechanical way. Today's HMDs use LCD or OLED displays and magnetic head tracking technology instead of mechanical. HMDs provide a stereoscopic view through displaying slightly different images to each eye creating the illusion of 3D. Head movement tracking is another important factor in order to show different parts of the virtual world to the participant when he/she moves his head to another direction or navigates in the environment. Every time the user moves his/her head in the environment, the new image that should be seen by the user is computed again which requires a computational effort. Some HMDs or wearable glasses combine the real and computer generated view in front of the user's eye through reflecting the real world view with mirrors which is called augmented reality. In projection-based systems, computer-generated 3D scene is projected on one or more walls and floor from different perspectives using stereoscopy. In [76] a number of projection-based displays are collected according to the number of walls used such as Powerwall (single wall), Immersion Square (three-wall), CAVE (four-wall), CABIN (five-wall) and six-sided displays such as IVY, COSMOS and ALICE. As the number of walls increases a more enclosed, realistic environment is created. However it brings some disadvantages such as projector placement and head tracking [76]. Factors such as resolution, field of view and frame rate are important properties of display devices influencing the quality of immersion. Besides the display hardware, some other efforts can create the illusion of depth on a single image such as perspective, occlusion, shadows, shading and texture [160]. Thalmann et. al [58] also mentions about the importance of realistic shape modeling and realistic illumination modeling for image level realism. In order to achieve believability, expected effects of display hardware should be consistent with the effects created on the static image. For example, when the user moves his head the effect of illumination will probably change and needs to be updated according to the current situation.

Auditory Sensory Channel. Auralization can be defined as the illusion of hearing a voice coming from a sound source that doesn't really exist with same acoustic conditions if it would be real. For example think that you are immersed into the virtual and you are together with a virtual human in a long corridor. Your virtual friend realizes you and start walking towards you. For full immersion, it is necessary to hear the footsteps considering the acoustic conditions of the room and your relative positions. As he/she comes close to you, you will hear the footsteps more clearly and more louder. Auralization is the rendering of sound through physical and mathematical modeling. A good overview can be found in [51] and [89]. It is used in many applications such as the simulation of architectural acoustics of buildings such as concert halls and theatres either existing or non-existing. Telepresence is another area where users participate in a remote physical environment just like they are really there and experience the sound in this environment. We are interested in auralization in virtual environments in order to increase feeling of presence and this notion is mentioned in several studies such as [176], [188] and [152]. In virtual environments, examination of the sound effects created by moving objects is rather an important topic considering the animation factor. Three principal components of auralization systems are sound source, medium and receiver [25]. The spatial location/orientation and physical properties of the sound source are factors to be considered in source modeling. Medium modifies the sound coming from the sound source through attenuation, reverberation, and Doppler affect. Sound propagates in the air and interferes with objects in the environment. Propagation of sound in the air results an attenuation of its intensity which is based on the distance between the sound source and the listener. Reverberation is the reflection of sound to listener after interfering with several objects in the environment and this is again related with distance since sound originating from a distant source will interfere with more objects and be effected more from occlusion and obstruction. Doppler effect is another phenomenon related with moving sound sources. Sound waves coming from a stationary source propagate uniformly however this is not the case for moving sound sources. Sound waves in the direction of movement become closer to each other and the ones at the opposite direction are more spread. In other words the frequency of the sound will be higher when the source is close and it will decay as the source moves away from the receiver. Third component in auralization is receiver. Position and direction of receiver are important properties as they are for the sound source. For virtual environments, binaural human-like hearing should be modeled since there is difference between the perception of two ears both in terms of intensity and time. This property is especially important in hearing sound sources in the horizontal plane and to distinguish between different sound sources. In [188], two approaches are mentioned for acoustic modeling of sound. Physical approach considers the physical configuration of the room and perceptual approach is based on imitating the perceived audible impression. Physical approach uses parameters such as reflection of surfaces and direction of sound waves and well adapted to virtual environments through changing these parameters in real time. They use some geometric models such as ray-tracing and image-source method and mathematical models such as finite element method and boundary element method which are also called wave-based models. Image-source method is most widely preferred in real-time applications. The important point for creating sound effects in real-time is parameterization. Image-sources are calculated according to the parameters such as position/orientation of source/listener and materials of the surfaces and interpolation is applied to these parameters in real time. An overview of these modeling techniques can be found in [188] and [152]. To complete, we will mention about three sound rendering techniques in the literature: Head-Related-Transfer-Function (HRTF) and Wave-field Synthesis (WFS) [25]. HRTF considers the rendering of sound according to one listening point and models the binaural hearing of human. Binaural headphones are used with sound filters modeling HRTF. WFS don't depend on the position and orientation of source and listener and is more suitable for multi-user environments. This kind of systems is mostly preferred in Cave-like systems where multiple users can experience the virtual world simultaneously. An example can be found in [18] where a multi-channel speaker system is described in the multi-screen immersive projection system CABIN.

Haptic Sensory Channel. Visual and auditory immersion is not enough for full immersion into the virtual environment. When trying to interact with a virtual object, if you cannot grasp the object after reaching and your hand passes through it, this will immediately decrease the feeling of presence and remind you that the environment is not real [129]. The term haptic is related with creating human sense of touch in a virtual world [58]. Haptic interaction between user and virtual environment is provided through haptic interface devices. Haptic devices are mechanical interfaces which convert signals from the computer into a form that user can perceive and also converts the signals from user to a computer understandable form. Haptic devices are different from visual and auditory devices since they realize a bidirectional channel between user and virtual [151]. While interacting with an object we can either feel a force response or feel the heat, pressure and texture of the object. Haptic devices are devided into two sub-groups: force (kinesthetic) feedback devices that employs our muscles, tendons, joints and tactile (cutaneous) devices that effects the receptors embedded in the skin [105]. In [58] three classes of haptic devices are defined: arm-like devices, exoskelatons and tactile displays. Arm-like devices such as PHANTOM are small desk-grounded devices that allows a single point contact with virtual object using a pen-like tool. Exoskelatons that a person wears on the hand or arm present multiple-degrees of freedom through providing force-feedback for each finger (e.g CyberGrasp). Inertia is a problem of these devices since the weight of the device effects the results of the computations. Tactile displays such as CyberTouch give tactile feedback rather than force feedback. Haptic rendering is defined as the process of generating forces as a result of the interaction with virtual objects and haptic rendering pipeline consists of two steps: the computation of the position and orientation of the virtual object grasped by the user and computation of the force and tourque feedback [105]. These two steps are also named as collision detection and collision response. Most of the haptic rendering algorithms are based on one point interaction with the object and allows three degrees-of-freedom since a point in 3D has three degrees-of-freedom. However this is not enough to model the interaction between objects such as when we are eating with a fork or writing with a pen [105]. For example, McNeely et al.[195] implemented six degrees-of-freedom using voxels. Haptic deformation of objects is another point to consider in the area of haptics and it is mostly used in haptic sculpting [77]. Haptics is used in many areas such as medical training, vehicle training, equipment maintenance training, cultural applications, textile and games. One interesting example of haptics is a haptic museum [113] in order to let the people touch valuable artifacts that are not allowed to touch in real or exhibiting all the items at the same time eliminating the space constraints. In [205], an immersive virtual reality environment for performing assembly and maintenance simulation is described. At MIRALab, a study [14] is developed in the context of HAPTEX project that presents visual and haptic rendering techniques for multimodal perception of textile. The aim is to create the feeling of touch on a cloth surface. Haptics is also used for artistic purposes such as haptic painting with virtual brushes [24].

However most of the studies involve the interaction between a person and virtual objects and there are rather few studies that investigates the human-virtual human interaction. This is because of the reason that today's haptic interface devices are limited in generating human like touching response. Bailenson [81] mentions about the notion of visual interpersonal touch in a collaborative virtual environment through the use of haptic devices. The aim of this study is to emphasize the role of touch in realistic social interaction e.g. handshaking of virtual characters. The result of the study show that users prefer to apply less force to faces when compared with liveless objects and less force to woman when compared to men. Another study by the same group investigates the role of touch in emotional behavior [48]. For example, touch can convey the feeling of trust e.g. a trustful or cheerful handshake. Haptic force feedback is also an important factor to increase presence while performing a collaborative task where participants simultaneously feel and manipulate objects such as carrying a heavy object together ([61] [140]).

2.5.2 Social Presence

Social presence is defined as the feeling that another intelligent character exists in the environment. With intelligence we mean the human-like intelligence and this requires the simulation of social interaction between user and virtual human. Social interaction is a very broad term containing the notion of face-to-face communication using several modalities such as speech, facial expressions and gestures. Head node and gaze are also used during conversation meaning that the virtual character understands and give respond to what the user is saying. Virtual characters do not only need to use their bodies in a human-like way but also should have distinct personalities and backgrounds as well as emotional behavior. The role of non-verbal modalities and emotion in communication is explained in detail in the interaction and personification sections. However, most of the studies we mentioned before focus on the realistic generation of gestures, facial expressions, speech or we examined them looking at this perspective. One point we did not consider in face-to-face communication is the awareness of virtual human. Recognition of the user and real environment is necessary to realize the user behavior and some information of the environment is required to make decisions related with real environment. Some of the ECAs we mentioned before consider the recognition of human gestures, facial expressions and speech. For example REA (Real Estate Agent) recognizes the user when he/she walks towards the screen and signals her awareness of the user by posture changes. MACK [182] has the ability to reference a shared physical space with the user in order to show the direction of places in the campus. Cassell et. al. [182] names this property as the immersion of virtual to real and looks at immersion from the other side. In [133] social awareness is mentioned as an important component of presence and believability. In this study, a storytelling drama on the site of ancient Pompeii is developed considering the construction of ancient frescos-paintings through the real time revival of fauna and flora and virtual animated characters with dramaturgical behaviors in an immersive AR environment. To increase the feeling of presence, once the real people are immersed into ancient Pompeii and see the virtual characters, virtual characters as well will look at the user, become closer and start talking. Effective interaction between real and virtual requires multimodal input recognition which is an area that concerns many fields such as human computer interaction, robotics, virtual reality etc. In this section, we will mention about recognition of speech, facial expressions, gestures and emotion recognition.

Recognizing the Real Human. For a natural interaction and control of the virtual environment users should speak with their natural voice in real time. For believable interaction, virtual human should recognize who is talking in a multiple user environment and understand the semantical and emotional content in the speech in order generate meaningful responds. Another reason for natural speech in virtual environments is controlling the environment without using a text-based device and freeing the hands in order to use haptic devices such as hand gloves. One also should distinguish between the terms automatic speech recognition and automatic speech understanding. Speech understanding is related with natural language understanding techniques. Speech recognition is related with converting acoustic speech signals to text. An overview of speech recognition techniques can be found in [202]. They mention about various dimensions in speech recognition such as speaker dependency, complexity of vocabulary etc. Speaker dependent systems are developed for use with single speaker. They are easier to develop and give accurate results but they are not flexible as speaker independent systems. Speaker independent systems are expensive and difficult to develop. There are also speaker adaptive systems used for a fixed small group of speakers where the system adapts itself to the characteristics of a new speaker. Small vocabulary systems are easier to develop but the requirement of small or large vocabulary changes according to the aim of the application. There are also systems that can understand continuous speech and systems where each word is given as input separately since detection of word boundaries is a difficult task. Slow or fast speech and co-articulation also affects the extraction of words and phonemes. Continuous and spontaneous speech recognition is necessary in virtual environments where it is possible also to understand laughter and coughing. Applying some semantic and syntactic language constraints by ignoring incomplete sentences can help to understand spontaneous speech. [202] also mentions that beside the language constraints task constraints can also be applied that ignores the sentences that are not related with a specific topic. Recognition in noisy environments and room acoustics can also affect the accuracy of speech recognition. Recently, speech recognition goes into the direction of catching emotional cues in the speech. For example, in [132] a large-scale data mining experiment about the automatic recognition of basic emotions in everyday short utterances is presented. An overview of emotional speech recognition techniques can be found in [184] which examines emotional speech recognition keeping in mind three goals: an up-to-date record of available emotional speech data, acoustic features used in emotion recognition (e.g. pitch, mel-frequency cepstral coefficients) and techniques used to classify speech into emotional states (e.g. hidden markov models, neural networks etc.).

Recognition of emotions and intentions from facial expressions is a rather older area when compared to emotional speech recognition. Facial expressions give clues about the cognitive and affective state of a person. According to [45] facial expression recognition is concerned with the deformations of facial components and their spatial relations as well as color changes on the skin. Steps in face recognition can be grouped into three: preprocessing, facial expression extraction and facial expression classification. In order to determine the exact place of the face in a given image, face detection is applied except the systems that have an automatic setup for image acquisition [111]. Some pre-processing is applied for removing noise and tracking the face parts before feature extraction [45]. Feature extraction converts the pixel information into high-level, meaningful features such as shape, motion, spatial configuration of facial components and color through reducing the dimensionality of data [111]. Facial expression classification is the mapping of a feature or a set of features to predefined labels of emotions [45]. Probably, the most well known system for facial expressions is the Facial Action Coding System of Ekman. Gaze recognition is also another important component for conveying emotions and impressions and we consider it here as a part of the facial expression recognition. Eye tracking technology makes it available to track the direction of the eye gaze through laser or cameras with devices worn on the head.

Gesture recognition is a new area when compared to facial expression recognition however it is an important component in recognition since it complements both speech and facial expressions. For example, gestures show how to do a task or convey an information about the current cognitive/emotional state. Whole body posture also gives information about the current affective and cognitive state. In [119], it is mentioned that gesture recognition does not only mean tracking of human movement but also the interpretation of that movement semantically. Interpretation can vary in a range of large and small scale motions e.g. tracking of small-scale movements, tracking the subject as a single object or as an articulated kinematic structure [119]. Gesture recognition is necessary to interpret this extra information and a virtual character that can recognize and interpret this information will look more intelligent. In [158], it is mentioned that if the user can not get the desired response to his/her gestural movements, this will decrease the quality of interaction and cause a break in the experience of presence. However, it is a difficult task to capture spontaneous gestures in virtual environments. Gesture-driven interfaces in virtual environments especially use emblematic and pantomime gestures which are more clear in their meaning [111]. Emblematic gestures are culture specific and are learned in time and have predefined, common meanings. Pantomime gestures describe a task or object and easier to interpret. Traditionally, gesture recognition in virtual environments was used for object manipulation by many researchers such as [26], [162] and [83]. A first example that uses deictic (pointing) gestures is "Put That There" [31], where user points at a place on the screen and explains a command such as

”Draw a blue square on the screen”. A more recent example of deictic gesture application can be found in [170]. In this system, user navigates in the environment by pointing in one direction on the screen. Gesture recognition systems designed for disabled people also exist that understands sign languages such as ”American Sign Language” [70]. There are different techniques and algorithms for gesture recognition. Turk [179] defines two groups of gesture recognition: tracker-based gesture recognition and vision-based gesture recognition. The former group contains data gloves and body suits (motion capture devices). Data gloves are mostly used for grasping operations, and with body suits it becomes available to track the general body movements with a few markers attached on the body. The disadvantage of these devices is that they can be difficult to setup and need calibration. They are also expensive devices and users have to wear these cumbersome devices. Vision-based systems eliminate the problems of using cumbersome and expensive devices. Gestures are captured with one or more cameras and features in the images are extracted in order to interpret the meaning of body movements. However, they have their own problems, for example it is not possible to track the full body from every direction because of occlusions. Algorithmic techniques such as feature extraction, template matching, statistical methods (e.g. principal component analysis) and machine learning techniques (e.g. neural networks and hidden markov models) are applied to interpret the captured data. An overview of these methods and how they are used in gesture recognition can be found in [84]. Affective posture recognition is an area that aims to map a specific posture to an emotion label similar to facial expression recognition. However different from facial expressions there is no system that relates low-level movements of the body to affective states [6]. In [6] an affective posture recognition system is introduced that can learn to label affective states with the help of explicit feedback sent to the system to correct the label.

2.6 Conclusion

In this paper, we have presented a discussion about the various aspects of designing and modeling virtual human and how intelligent decision technologies can be used in many stages of virtual human architectures. Autonomy, interaction, personification and presence are four important aspects considering both mental state of a virtual human and its interaction with virtual environment and real people. The area has its basis in different disciplines so that successful results in this field is in majority due to good collobaration between different communities. With the many advances in intelligent decision technologies in recent years, these techniques are becoming more important to model the mind of the virtual human as well as analyzing the captured data from the real human and modeling the interaction between virtual and real. Utilizing these techniques for the creation of virtual human will surely light the way for designing more realistic virtual characters in addition to the improvements in computer graphics.

Acknowledgement

This research is reported in the International Journal of Intelligent Decision Technologies, IOS Press, Vol. 1, No. 1-2, pp. 3-15, 2007.

References

1. Wikipedia-action-selection (2006), `http://en.wikipedia.org/wiki/Actionselection`
2. Egges, A., Magnenat-Thalmann, N.: Emotional communicative body animation for multiple characters. In: V-Crowds 2005, Lausanne, Switzerland, pp. 31–40 (2005)
3. Monzani, J., Caicedo, A., Thalmann, D.: Integrating behavioural animation techniques. Computer Graphics Forum 20(3) (2001)
4. Romano, D., Sheppard, G., Hall, J., Miller, A., Ma, Z.: Basic: a believable, adaptable socially intelligent character for social presence. In: PRESENCE 2005, The 8th Annual International Workshop on Presence (2005)
5. Magnenat-Thalmann, N., Egges, A., Molet, T.: Personalised real-time idle motion synthesis. In: Pacific Graphics 2004, Seoul, Korea, pp. 121–130 (2004)
6. Bianchi-Berthouze, N., Kleinsmith, A., Fushimi, T.: An incremental and interactive affective posture recognition system. In: International Workshop on Adapting the Interaction Style to Affective Factors (2005)
7. Collins, A., Ortony, A., Clore, G.L.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
8. Sobral, D., Paiva, A., Dias, J., Aylett, R.: Caring for agents and agents that care: building empathic relations with synthetic agents. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (2004)
9. Zambetta, F., Paradiso, A., Abbattista, F.: Fanky: A tool for animating faces of 3d agents. In: de Antonio, A., Aylett, R.S., Ballin, D. (eds.) IVA 2001. LNCS (LNAI), vol. 2190, pp. 242–244. Springer, Heidelberg (2001)
10. Wyvill, G., Pearce, A., Wyvill, B., Hill, D.: Speech and expression: A computer solution to face animation. In: Graphics Interface 1986, pp. 136–140 (1986)
11. Karpouzis, K., Raouzaiou, A., Kollias, S.: Emotion synthesis in virtual environments. In: 6th International Conference on Enterprise Information Systems (2004)
12. Webpage AIML (September 2006), `http://www.alicebot.org`
13. Bindiganavale, R., Schuler, W., Allbeck, J.M., Badler, N.I., Joshi, A.K., Palmer, M.: Dynamically altering agent behaviors using natural language instructions. In: Proceedings of the fourth international conference on Autonomous agents, pp. 293–300 (2000)
14. Volino, P., Davy, P., Bonanni, U., Magnenat-Thalmann, N., Böttcher, G., Allerkamp, D., Wolter, F.-E.: From measured physical parameters to the haptic feeling of fabric. In: Proceedings of the HAPTEX 2005 Workshop on Haptic and Tactile Perception of Deformable Objects, pp. 17–29 (2005)
15. Arafa, Y., Kamyab, K., Mamdani, E.: Character animation scripting languages: a comparison. In: AAMAS 2003: Proceedings of the second international joint conference on Autonomous agents and multiagent systems, pp. 920–921 (2003)
16. Arbantes, G., Pereira, F.: Mpeg-4 facial animation technology: Survey, implementation, and results. IEEE Transactions on Circuits and Systems for Video Technology 9(2), 290–305 (1999)

17. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge (1976)
18. Ogi, T., Kayahara, T., Kato, M., Asayama, H., Hirose, M.: Immersive sound field simulation in multi-screen projection displays. In: Proceedings of the workshop on Virtual environments, pp. 135–142 (2003)
19. Churcher, G.E., Atwell, E.S., Souter, C.: Dialogue management systems: A survey and overview. Technical report, University of Leeds (1997)
20. Aylett, R., Luck, M.: Applying artificial intelligence to virtual reality: Intelligent virtual environments. Applied Artificial Intelligence 14(1), 3–32 (2000)
21. Pelachaud, C., Hartmann, B., Mancini, M.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. In: Gesture Workshop. LNCS (LNAI), pp. 188–199. Springer, Heidelberg (2005)
22. Lee, S.P., Badler, J.B., Badler, N.I.: Eyes alive. ACM Transactions on Graphics 21(3), 637–644 (2002)
23. Bartneck, C.: Integrating the occ model of emotions in embodied characters. In: Proceedings of the Workshop on Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges, Melbourne (2002)
24. Baxter, W., Scheib, V., Lin, M., Manocha, D.: Dab: Interactive haptic painting with 3d virtual brushes. In: Proceedings of ACM Siggraph, pp. 461–468 (2001)
25. Begault, D.: 3-D Sound for Virtual Reality and Multimedia. Academic Press, Cambridge (1994)
26. Bergamasco, M.: Manipulation and exploration of virtual objects. In: Magnenat Thalmann, N., Thalmann, D. (eds.) Artificial Life and Virtual Reality. John Wiley, Chichester (1994)
27. Bickmore, T., Cassell, J.: Social dialogue with embodied conversational agents. In: van Kuppevelt, J., Dybkjaer, L., Bernsen, N. (eds.) Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems. Kluwer Academic, New York (2005)
28. Biocca, F.: The cyborg's dilemma: Progressive embodiment in virtual environments. Journal of Computer-Mediated Communication (1997)
29. Taylor, P., Black, A., Caley, R.: The architecture of the festival speech synthesis system. In: 3rd ESCA Workshop on Speech Synthesis, Australia, pp. 147–151 (1998)
30. Ardissono, L., Boella, G., Lesmo, L.: An agent architecture for nl dialogmodeling. In: Proc. Second Workshop on Human-Computer Conversation (1998)
31. Bolt, R.A.: Put-that-there: Voice and gesture at the graphics interface. In: Proceedings of SIGGRAPH, pp. 262–270 (1980)
32. Bordeux, C., Boulic, R., Thalmann, D.: An efficient and flexible perception pipeline for autonomous agents. In: Computer Graphics Forum (Eurographics 1999), vol. 18(3), pp. 23–30 (1999)
33. Bos, J., Oka, T.: Building spoken dialogue systems for believable characters. In: Proceedings of the seventh workshop on the semantics and pragmatics of dialogue, Diabruck (2003)
34. Brand, M.: Voice puppetry. In: Proceedings of SIGGRAPH 1999, pp. 21–28 (1999)
35. Brand, M., Hertzmann, A.: Style machines. In: SIGGRAPH 2000, pp. 399–407. ACM Press, New York (2000)
36. Bratman, N.: Intention, plans, and practical reason. Harvard University Press, Cambridge (1987)
37. Sadek, M.D., Bretier, P., Panaget, F.: Artimis: Natural dialogue meets rational agency. In: Pollack, M.E. (ed.) Proceedings 15th International Joint Conference on Artificial Intelligence, pp. 1030–1035 (1997)

38. Brooks, R.A.: How to build complete creatures rather than isolated cognitive simulators. In: VanLehn, K. (ed.) Architectures for Intelligence, pp. 225–239. Lawrence Erlbaum Assosiates, Mahwah (2001)
39. Lombard, M., Reich, R.D., Grabe, M.E., Bracken, C., Ditton, T.B.: Presence and television: The role of screen size. Human Communication Research 26(1), 75–98 (2000)
40. Cañamero, L.: Emotions for activity selection. Dept. of Computer Science Technical Report DAIMI PB 545, University of Aarhus, Denmark (2000)
41. Canamero, D.: Designing emotions for activity selection. Department of Computer Science, University of Aarhus (2000)
42. Gratch, J., Rickel, J., Andre, E., Badler, N., Cassell, J., Petajan, E.: Creating interactive virtual humans: Some assembly required. IEEE Intelligent Systems 17(4), 54–63 (2002)
43. Chang, P.H.-M., Chien, Y.-H., Kao, E.C.-C., Soo, V.-W.: A knowledge-based scenario framework to support intelligent planning characters. LNCS, pp. 134–145. Springer, Heidelberg (2005)
44. Chopra-Khullar, S., Badler, N.I.: Where to look? automating attending behaviors of virtual human characters. In: AGENTS 1999: Proceedings of the third annual conference on Autonomous Agents, pp. 16–23 (1999)
45. Chibelushi, F.B.C.C.: Facial expression recognition: A brief tutorial overview. School of Computing, Staffordshire University (2002)
46. Cohen, M., Massaro, D.: Modelling coarticulation in synthetic visual speech. In: Thalmann, N.M., Thalmann, D. (eds.) Models and Techniques in Computer Animation, pp. 139–156 (1993)
47. Cornelius, R.: The science of emotion. Prentice Hall, New Jersey (1996)
48. Bailenson, J.N., Yee, N., Brave, S., Merget, D., Koslow, D.: Virtual interpersonal touch: Expressing and recognizing emotions through haptic devices. Human-Computer Interaction (in press, 2006)
49. Pearce, A., Hill, D., Wyvill, B.: Animating speech: an automated approach using speech synthesis by rules. The Visual Computer 3, 277–289 (1988)
50. Downie, M., Isla, D., Burke, R., Blumberg, B.: Simhuman: A platform for real-time virtual agents with planning capabilities. In: Proceedings of IJCAI, pp. 1051–1058 (2001)
51. Kleiner, M., Dalenback, B.-I., Svensson, P.: Auralization - an overview. Journal of the Audio Engineering Society 41(11) (1993)
52. de Gelder, B.: Towards the neurobiology of emotional body language. Nature Reviews Neuroscience 7(3), 242–249 (2006)
53. de Sevin, E., Thalmann, D.: An affective model of action selection for virtual humans. In: Proceedings of Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action symposium at the Artificial Intelligence and Social Behaviors Conference, pp. 293–297 (2005)
54. Dinerstein, J., Egbert, P.K.: Fast multi-level adaptation for interactive autonomous characters. ACM Trans. Graph 24(2), 262–288 (2005)
55. Traum, D., Larsson, S.: The information state approach to dialogue management. In: Smith, Kuppevelt (eds.) Current and New Directions in Discourse and Dialogue, pp. 325–353. Kluwer Academic Publishers, Dordrecht (2000)
56. Nakamura, S., Yamamato, E., Shikano, K.: Lip movement synthesis from speech based on hidden markov models. Speech Communication 26, 105–115 (1998)
57. Egges, A.: Real-time animation of interactive virtual characters. PhD thesis, MIRALab, University of Geneva (2006)

58. Magnenat Thalmann, N., Kim, H., Egges, A., Garchery, S.: Believability and interaction in virtual worlds. In: International Multi-Media Modelling Conference. IEEE Computer Society Press, Los Alamitos (2005)
59. Ekman, P.: Emotion in the Human Face. Cambridge University Press, New York (1982)
60. Brom, C., Lukavský, J., Šerý, O., Poch, T., Šafrata, P.: Affordances and level-of-detail ai for virtual humans. In: Proceedings of Game Set and Match 2 The Netherlands, pp. 134–145 (2006)
61. Eva-Lotta, Rassmus, K., Calle: Supporting presence in collaborative environments by haptic force feedback. ACM Transactions on Computer-Human Interaction 7(4), 461–476 (2000)
62. Ferguson, I.A.: TouringMachines: An Architecture for Dynamic, Rational, Mobile Agents. PhD thesis, Clare Hall, University of Cambridge (1992)
63. Gockley, R., Forlizzi, J., Simmons, R.: Interactions with a moody robot. In: Proceedings of Human-Robot Interaction, pp. 186–193 (2006)
64. Friesen, E.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)
65. Gadanho, S.: Learning behavior-selection by emotions and cognition in a multi-goal robot task. Journal of Machine Learning 4, 385–412 (2003)
66. Gebhard, P.: Alma - a layered model of affect. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi Agent Systems, pp. 29–36 (2005)
67. Gillies, M., Ballin, D.: Integrating autonomous behavior and user control for believable agents. In: AAMAS 2004: Proceedings of the first international joint conference on Autonomous agents and multiagent systems (2004)
68. Gillies, M.F.P., Dodgson, N.A.: Eye movements and attention for behavioural animation. The Journal of Visualization and Computer Animation 13(5), 287–300 (2002)
69. Grandstrom, B.: Multi-modal speech synthesis with applications. In: 3rd International School on Neural Nets, pp. 136–140 (1999)
70. McGuire, R.M., Hernandez-Rebollar, J., Starner, T., Henderson, V., Brashear, H., Ross, D.S.: Towards a one-way american sign language translator. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (2004)
71. Hampson, S.: State of the art: Personality. The Psychologist 12(6), 284–290 (1999)
72. Heeter, C.: Being there: The subjective experience of presence. Presence: Teleoperators and Virtual Environments 1(2), 262–271 (1992)
73. Herrero, P., de Antonio, A.: Introducing human-like hearing perception in intelligent virtual agents. In: AAMAS 2003: Proceedings of the second international joint conference on Autonomous agents and multiagent systems, pp. 733–740 (2003)
74. Heylen, D.: A closer look at gaze. In: Pelachaud, C., André, E., Kopp, S., Ruttkay, Z.M. (eds.) Creating Bonds with Embodied Conversational Agents: AAMAS Workshop (2005)
75. Kim, Y., Hill, R.W., Traum, D.R.: A computational model of dynamic perceptual attention for virtual humans. In: 14th Conference on Behavior Representation in Modeling and Simulation (2005)
76. Hogue, A., Robinson, M., Jenkin, M.R., Allison, R.S.: A vision-based head tracking system for fully immersive displays. In: Deisinger, J., Kunz, A. (eds.) 7th International Immersive Projection Technologies Workshop in conjunction with the 9th Eurographics Workshop on Virtual Environments (2003)

77. Hui, C., Hanqiu, S., Xiaogang, J.: Interactive haptic deformation of dynamic soft objects. In: Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications, pp. 255–261 (2006)
78. de Rosis, F., Poggi, I., Pelachaud, C.: Eye communication in a conversational 3d synthetic agent. André, E. (ed.): Special Issue of Artificial Intelligence Communications, The European Journal on Artificial Intelligence 13(3), 169–181 (2000)
79. Spindel, M.S., Roseman, I., Jose, P.E.: Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. Personality and Social Psychology 59(5), 899–913 (1990)
80. Isla, D.: Handling complexity in halo 2. In: VanLehn, K. (ed.) Architectures for Intelligence, Accesses September 2006. Lawrence Erlbaum Assosiates, Mahwah (2006)
81. Bailenson, J., Yee, N.: Virtual interpersonal touch: Haptic interaction and copresence in collaborative virtual environments. International Journal of Multimedia Tools and Applications (in press, 2006)
82. Champandard, A.J.: AI Game Development: Synthetic Creatures with learning and Reactive Behaviors. New Riders (2003)
83. James, D., Shah, M.: Gesture recognition. Technical Report, Department of Computer Science, University of Central Florida, CS-TR-93-11 (1993)
84. Joseph, L.: A survey of hand posture and gesture recognition techniques and technology. Technical Report CS-99-11, Brown University, Department of Computer Science (1999)
85. Kasper, M., Fricke, G., Steuernagel, K., Von Puttkamer, E.: A behavior-based mobile robot architecture for learning from demonstration. Robotics and Autonomous Systems 34, 153–164 (2001)
86. Kallmann, M.: Object interaction in real-time virtual environments. PhD Thesis. EPFL, Switzerland (2002)
87. Magnenat-Thalmann, N., Kalra, P.: Modeling of vascular expression in facial animation. In: Proc. CA 1994, pp. 50–58. IEEE Computer Society Press, Los Alamitos (1994)
88. Khirsagar, S.: Facial Communication. MIRALab, University of Geneva (2003)
89. Kleiner, M.: Auralization: a dsp approach. In: Sound and Video Contractor (1992)
90. Tepper, P., Kopp, S., Cassell, J.: Content in context: Generating language and iconic gesture without a gestionary. In: Proceedings of the Workshop on Balanced Perception and Action in ECAs at AAMAS 2004 (2004)
91. Kopp, S., Wachsmuth, I.: A knowledge-based approach for lifelike gesture animation. In: Horn, W. (ed.) ECAI 2000 (2000)
92. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. IEEE Intelligent Systems 15(1), 39–52 (2004)
93. Egges, A., Kshirsagar, S., Magnenat-Thalmann, N.: Generic personality and emotion simulation for conversational agents. Computer Animation and Virtual Worlds 15(1), 1–13 (2004)
94. Kuffner, J.J., Latombe, J.C.: Fast synthetic vision, memory, and learning models for virtual humans. In: Proceedings of CA 1999: IEEE International Conference on Computer Animation (1999)
95. Gleicher, M., Kovar, L., Pighin, F.: Motion graphs. ACM Transactions on Graphics 21(3), 473–482 (2002)
96. Laban, R., Lawrence, F.C.: Embodied Conversational Agents. Plays, Inc., Boston (1974)

97. Lamarche, F., Donikian, S.: Automatic orchestration of behaviours through the management of resources and priority levels. In: AAMAS 2002: Proceedings of the first international joint conference on Autonomous agents and multiagent systems (2002)
98. Lavagetto, F., Pockaj, R.: The facial animation engine: towards a high-level interface for the design of mpeg-4 compliant animated faces. IEEE Transactions on Circuits and Systems for Video Technology 9(2), 277–289 (1999)
99. Lazharus, R.: Emotion and Adaptation. Oxford University Press, Oxford (1991)
100. Lee, K.M., Nass, C.: Proceedings of the sigchi conference on human factors in computing systems. In: Proceedings of the SIGCHI conference on Human factors in computing systems, vol. 33, pp. 289–296 (2003)
101. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lee, A., Bregler, C.: Speaking with hands: Creating animated conversational characters from recordings of human performance. ACM Transactions on Graphics 23(3), 506–513 (2004)
102. Lehnert, H.: Fundamentlas of auditory virtual environment. In: Thalmann, N.M., Thalmann, D. (eds.) Artificial Life and Virtual Reality. John Wiley and Sons Ltd., Chichester (1994)
103. Klein, J.B.E., Lemon, O., Oka, T.: Dipper: Description and formalisation of an information-state update dialogue system architecture. In: 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan (2003)
104. Lewis, J.: Automated lip-sync: Background and techniques. The Journal of Visualization and Computer Animation 2, 118–122 (1991)
105. Otaduy, M.A., Lin, M.C.: Introduction to haptic rendering. In: ACM Sigraph Course Notes (2005)
106. Becker, C., Nakasone, A., Prendinger, H., Ishizuka, M., Wachsmuth, I.: Physiologically interactive gaming with the 3d agent max. In: International Workshop on Conversational Informatics, pp. 37–42 (2005)
107. Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M.P.: Integrated learning for interactive synthetic characters. In: Proceedings of SIGGRAPH 2002, pp. 417–426 (2002)
108. Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., Hagita, N.: Messages embedded in gaze of interface agents - impression management with agent's gaze. In: Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves (2002)
109. Martin, J.C., Pelachaud, C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M.: Levels of representation in the annotation of emotion for the specification of expressivity in ecas. In: Proceedings of Intelligent Virtual Agents (2005)
110. Silverman, B.G., Johns, M.: How emotion and personality effect the utility of alternative decisions: A terrorist target selection case study. In: Proceedings of the 10th Conference on Computer Generated Forces and Behavioral Representation (2001)
111. Rothkrantz, L.J.M., Pantic, M.: Automatic analysis of facial expressions: the state of the art. IEEE Transactions Pattern Analysis Machine Intelligence 22(12), 1424–1445 (2000)
112. Renault, O., Magnenat-Thalmann, N., Thalmann, D.: A vision-based approach to behavioural animation. Journal of Visualization and Computer Animation 1(1), 18–21 (1990)
113. Shahabi, C., McLaughlin, M.L., Sukhatme, G.: Haptic museum. Technical Report, Information Laboratory, University of Southern California (2000)

114. Marsella, S., Gratch, J.: A step toward irrationality: using emotion to change belief. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 334–341 (2002)
115. Marsella, S., Gratch, J.: Modeling coping behavior in virtual humans: Don't worry, be happy. In: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (2003)
116. Mccrae, R.R., John, P.O.: An introduction to the five-factor model and its applications. Journal of Personality 60, 175–215 (1992)
117. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
118. McTear, M.: Spoken dialogue technology: Enabling the conversational user interface. ACM Computing Survey 34(1) (2002)
119. Vance, J.M., Jerney, M.M.: Gesture recognition in virtual environments: A review and framework for future development. Iowa State University Human Computer Interaction, Technical Report ISU-HCI-2005-01 (2005)
120. Morishima, S.: Real-time talking head driven by voice and its application to communication and entertainment. In: International Conference on Auditory Visual Speech Processing (1998)
121. Muller, J., Pischel, M.: The agent architecture interrap: Concept and application. Technical Report RR-93-26, DFKI Saarbrucken (1993)
122. Nareyek, A.: Languages for programming bdi-style agents: An overview. In: Proc. of the Second International Conference on Computers and Games (2000)
123. Nijholt, A.: Humor and embodied conversational agents. CTIT Technical Report series No. 03-03, University of Twente, ISSN 1381-3625 (2003)
124. Noh, J., Neumann, U.: A survey of facial modeling and animation techniques. USC Technical Report 99-705 (1998)
125. Badler, N.I., Palmer, O., Bindiganavale, R.: Fast multi-level adaptation for interactive autonomous characters. Communications of ACM 42(8), 65–73 (1999)
126. Forsyth, F., Arikan, O.: Interactive motion generation from examples. ACM Transactions on Graphics 21(3), 483–490 (2002)
127. Forsyth, F., Arikan, O., O'Brien, J.F.: Motion synthesis from annotations. ACM Transactions on Graphics 22(3), 402–408 (2003)
128. Burghouts, G.J., op den Akker, H.J.A., Heylen, D.K.J., Poel, M., Nijholt, A.: An action selection architecture for an emotional agent. In: Proceedings of 16th International FLAIRS Conference, pp. 293–297 (2003)
129. Insko, B.E., Meehan, M.J., Whitton, M.C., Brooks Jr., F.P.: Passive haptics significantly enhances virtual environments. Computer Science Technical Report, University of North Carolina (2001)
130. Magnenat-Thalmann, N., Kalra, P., Mangili, A., Thalmann, D.: Simulation of facial muscle actions based on rational free form deformation. In: Proceedings Eurographics 1992, pp. 59–69 (1992)
131. Bianchi-Berthouze, N., De Silva, P.R.: Modeling human affective postures: an information theoretic characterization of posture features. Journal of Visualization and Computer Animation 15(3-4), 269–276 (2004)
132. Oudeyer, P.-Y.: The production and recognition of emotions in speech: features and algorithms. International Journal in Human-Computer Studies, special issue on Affective Computing 59(1-2), 157–183 (2003)
133. Magnenat-Thalmann, N., Papagiannakis, G., Kim, H.: Believability and presence in mobile mixed reality environments. In: IEEE VR 2005 Workshop on Virtuality Structures (2005)

134. Parke, F.: Computer generated animation faces. In: ACM Annual Conference (1972)
135. Pelachaud, C.: Communication and Coarticulation in Facial Animation. University of Pennsylvania (1991)
136. Perlin, K., Goldberg, A.: Improv: A system for scripting interactive actors in virtual worlds. In: Computer Graphics. Annual Conference Series, vol. 30, pp. 205–216 (1996)
137. Peters, C., O'Sullivan, C.: Synthetic vision and memory for autonomous virtual humans. Computer Graphis Forum 21(4) (2002)
138. Peters, C., Sullivan, C.: Bottom-up visual attention for virtual human animation. In: CASA 2003: Proceedings of the 16th International Conference on Computer Animation and Social Agents (2003)
139. Picard, R.W.: Affective computing. MIT Press, Cambridge (1997)
140. Pinho, M., Bowman, D., Freitas, C.: Cooperative object manipulation in immersive virtual environments: Framework and techniques. In: Proceedings of ACM Virtual Reality Software and Technology, pp. 171–178 (2002)
141. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) Emotion: Theory, research, and experience, vol. 1, pp. 2–33 (1980)
142. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.: Embodied Conversational Agents. MIT Press, Cambridge (2000)
143. Soar Project. University of michigan. project homepage Accessed (September 2006), `http://sitemaker.umich.edu/soar`
144. Rabie, T.F., Terzopoulos, D.: Active perception in virtual humans. In: Vision Interface 2000 Montreal, Canada (2000)
145. Reynolds, C.W.: Flocks, herds, and schools: a distributed behavioral model. In: SIGGRAPH 1987 Conference Proceedings, vol. 21(4), pp. 25–34 (1987)
146. Rickel, J., Johnson, W.L.: Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. Applied Artificial Intelligence 13, 343–382 (1999)
147. Ruttkay, Z., Noot, H.: Variations in gesturing and speech by gestyle. International Journal of Human-Computer Studies, Special Issue on Subtle Expressivity for Characters and Robots 62(2), 211–229 (2005)
148. Becker, C., Kopp, S., Wachsmuth, I.: Simulating the emotion dynamics of a multimodal conversational agent. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 154–165. Springer, Heidelberg (2004)
149. Magnenat-Thalmann, N., Kshirsagar, S., Garchery, S.: Feature point based mesh deformation applied to mpeg-4 facial animation. In: Proceedings of the IFIP DEFORM 2000 Workshop and AVATARS 2000 Workshop on Deformable Avatars, pp. 23–34 (2000)
150. Pelachaud, C., Pasquariello, S.: Greta: A simple facial animation engine. In: 6th Online World Conference on Soft Computing in Industrial Appications (2001)
151. Salisbury, K., Conti, F., Barbagli, F.: Passive haptics significantly enhances virtual environments. Computer Science Technical Report, University of North Carolina 24(2), 24–32 (2004)
152. Savioja, L.: Modeling Techniques for Virtual Acoustics. PhD thesis, Helsinki University of Technology (1999)
153. Schroder, M.: Emotional speech synthesis: A review. In: Proceedings of Eurospeech 2001 (2001)

154. Catizone, R., Setzer, A., Wilks, Y.: State of the art in dialogue management (2002)
155. Sheridan, T.B.: Musings on telepresence and virtual presence. Presence-Connect 1, 120–126 (1992)
156. Badler, N., Bindiganavale, R., Bourne, J., Palmer, M., Shi, J., Schuler, W.: A parameterized action representation for virtual agents. In: Workshop on Embodied Conversational Characters (1998)
157. Slater, M.: A note on presence terminology. Presence-Connect 3 (2003)
158. Slater, M., Steed, A.: A virtual presence counter. Presence 9(5), 413–434 (2000)
159. Stead, L., Goulev, P., Evans, C., Mamdani, E.: Journal of ubiquitous computing. ACM Transactions on Graphics 8(3-4), 282–290 (2004)
160. Steed, A.: A survey of virtual reality literature. In Technical Report (1993)
161. Stronks, J.J.S.: Friendship relations with embodied conversational agents: Integrating social psychology in eca design. Master Thesis, Faculty of Computer Science, University of Twente (2002)
162. Sturman, D.J., Zeltzer, D.: A survey of glove-based input. IEEE Computer Graphics and Applications 14(1), 30–39 (1994)
163. Wardhani, A., Su, W., Pham, B.: High-level control posture of story characters based on personality and emotion. In: Pisan, Y. (ed.) Proceedings IE 2005, The Second Australasian Interactive entertainment conference, pp. 179–186 (2005)
164. Sutherland, I.: A head-mounted three-dimensional display. In: Proceedings of the Fall Joint Computer Conference, vol. 33, pp. 757–764 (1968)
165. Shin, S.Y., Kim, T., Park, S.: Rhythmic-motion synthesis based on motion-beat analysis. ACM Transactions on Graphics 22(3), 392–401 (2003)
166. Tang, T.W., Wan, R.: Integrated learning for interactive synthetic characters. In: WSCG - Short Papers, pp. 137–144 (2002)
167. Garcia-Rojas, A., Gutierrez, M., Thalmann, D., Vexo, F.: Multimodal authoring tool for populating a database of emotional reactive actions. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 206–217. Springer, Heidelberg (2006)
168. Noser, H., Renault, O., Thalmann, D., Magnenat Thalmann, N.: Navigation for digital actors based on synthetic vision memory and learning. Computers and Graphics 19(1), 7–19 (1995)
169. Blumberg, B., Todd, P., Maes, P.: No bad dogs: Ethological lessons for learning in hamsterdam. In: Proceedings of the 4th International Conference on the Simulation of Adaptive Behavior (1996)
170. Tollmar, K., Demirdjian, D., Darrell, T.: Navigating in virtual environments using a vision-based interface. In: Proceedings of the third Nordic conference on Human-computer interaction, pp. 113–120 (2004)
171. Tomlinson, B., Blumberg, B.: Alphawolf: Social learning, emotion and development in autonomous virtual agents. In: First GSFC/JPL Workshop on Radical Agent Concepts (2002)
172. Tomlinson, B.: From linear to interactive animation: how autonomous characters change the process and product of animating. ACM Computers in Entertainment 3(1) (2005)
173. Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., Swartout, W.: Toward a new generation of virtual humans for interactive experiences. IEEE Intelligent Systems 17(4), 32–38 (2002)
174. TRINDIKIT (September 2006), `http://www.ling.gu.se/projekt/trindi/trindikit/index.html`
175. Trung, H.B.: Multimodal dialogue management - state of the art. Human Media Interaction Department, University of Twente (2006)

176. Funkhouser, T., Tsingos, N., Jot, J.M.: Survey of methods for modeling sound propagation in interactive virtual environment systems. Presence and Teleoperation (2003)
177. Tu, X., Terzopoulos, D.: Artificial fishes: physics, locomotion, perception, behavior. In: Proceedings SIGGRAPH 1994, vol. 21(4), pp. 43–50 (1994)
178. Tu, X., Terzopoulos, D.: Perceptual modeling for the behavioral animation of fishes. In: Proc. Pacific Graphics 1994, pp. 165–178. World Scientific, Singapore (1994)
179. Turk, M.: Gesture recognition. In: Stanney, K. (ed.) Handbook of Virtual Environments: Design, Implementation and Applications, pp. 113–120. Lawrence Erlbaum Associates, Mahwah (2002)
180. Schuemie, M.J., Straaten, P., van der Krijn, M., van der Mast, C.: Research on presence in virtual reality: A survey. Presence-Connect 4(2), 183–202 (2001)
181. van Waveren, J.M.P.: The quake 3 arena bot. Master thesis. Faculty ITS, University of Technology Delft (2001)
182. Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, H.: MACK: Media lab autonomous conversational kiosk. In: Proceedings of Imagina 2002 (2002)
183. Veldhuis, J.: Expressing personality through head node and eye gaze. In: 5th Twente Student Conference on IT (2006)
184. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. Speech Communication 48 (2006)
185. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: Proceedings of SIGGRAPH (2001)
186. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: Proceedings of the CHI 1999 Conference (1999)
187. Mascardi, D.A.V., Demergasso, D.: Languages for programming bdi-style agents: An overview. In: WOA 2005 (2005)
188. Väänänen, R.: Parametrization, auralization, and authoring of room acoustics for virtual reality applications. In: Electrical and Communications Engineering (2003)
189. Vosinakis, S., Panayiotopoulos, T.: Simhuman: A platform for real-time virtual agents with planning capabilities. In: Proceedings of the 3rd International Workshop on Intelligent Virtual Agents (2001)
190. Wang, F., Mckenzie, E.: Virtual life in virtual environments. CSG report ECS-CSG-44-98, Institute for Computing Systems Architecture, University of Edinburgh (1998)
191. Waters, K., Levergood, T.: Decface: A system for synthetic face applications. Multimedia Tools and Applications 1(4), 349–366 (1988)
192. Weizenbaum, J.: Eliza-a computer program for the study of natural language communication between man and machine. ACM Computing Survey 9(1), 36–45 (1966)
193. Whissel, C.M.: The dictionary of affect in language. In: Plutchik, R., Kellerman, H. (eds.) Emotion: Theory, research, and experience. The measurement of emotions, vol. 4 (1980)
194. Wikipedia-Behaviour (accessed September 2006), `http://en.wikipedia.org/wiki/behaviour`
195. William, K.D., McNeely, W.A.: Six degrees-of-freedom haptic rendering using voxel sampling. In: ACM Sigraph Course Notes, pp. 401–408 (1999)

196. Wilson, I.: The artificial emotion engine, driving emotional behaviour. Artificial Intelligence and Interactive Entertainment (2000)
197. Wooldridge, M.: Reasoning about rational agents. MIT Press, Cambridge (2000)
198. Woolridge, M.: An Introduction to MultiAgent Systems. John Wiley and Sons, Chichester (2002)
199. Wang, T., Li, Y., Shum, H.Y.: Motion texture: a two-level statistical model for character motion synthesis. ACM Transactions on Graphics 21(3), 465–472 (2002)
200. Chang, Y., Ezzat, T.: Transferable videorealistic speech animation. In: ACM Siggraph/Eurographics Symposium on Computer Animation, Los Angeles, pp. 21–28 (2005)
201. Turunen, M., Wilks, Y., Catizone, R.,: Companions consortium: State of the art papers. 2 (2006)
202. Zhang, M.: Overview of speech recognition and related machine learning techniques. Technical Report, CS-TR-01/15 (2001)
203. Chi, D.M., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: Siggraph 2000, Computer Graphics Proceedings (2000)
204. Badler, N., Allbeck, J., Zhao, L., Byun, M.: Representing and parameterizing agent behaviors. In: Computer Animation, pp. 133–143 (2002)
205. Zhao, W., Madhavan, V.: Virtual assembly operations with grasp and verbal interaction. In: Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications, pp. 245–254 (2006)

3

A Comprehensive Context Model for Multi-party Interactions with Virtual Characters

Norbert Pfleger and Markus Löckelt

DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
{pfleger,loeckelt}@dfki.de
http://www.dfki.de/~pfleger

Abstract. Contextual information plays a crucial role in nearly every conversational setting. When people engage in conversations they rely on what has previously been uttered or done in various ways. Some nonverbal actions are ambiguous when viewed on their own. However, when viewed in their context of use their meaning is obvious. Autonomous virtual characters that perceive and react to events in conversations just like humans do also need a comprehensive representation of this contextual information. In this chapter we describe the design and implementation of a comprehensive context model for virtual characters.

3.1 Introduction

Contextual information influences the understanding and generation of communicational behavior and it is widely acknowledged that any multimodal dialogue system that deals with more or less natural input must incorporate contextual information [Bunt, 2000]. The goal of the work described here is to develop a generic and comprehensive context model that supports the integration of perceived monomodal events into a multimodal representation, the resolution of referring expressions, and the generation of reactive and deliberative actions. Besides the classical linguistic context provided by a discourse history, we have identified sets of physical and conversational context factors that a multimodal dialogue system aiming at real conversational interaction needs to incorporate. Using this extended notion of context, we aim at processing both natural nonverbal and verbal behavior in dyadic as well as multi-party conversations. The context-model has been incorporated into an integrated fusion and discourse processing engine (called FADE) that is employed in different multimodal dialogue systems. In this chapter we show how FADE has been employed in the VirtualHuman system, a conversational multi-party dialogue system where human users can interact with virtual characters.

This chapter is organized as follows: We start with a brief description of the VirtualHuman system. Then we outline the key aspects of our context model in Sec. 3.3 and give in Sec. 3.4 a brief overview of how we implemented this model within FADE. Sec. 3.5 provides a brief overview of how the action manager of the system processes the contextually disambiguated contributions.

N. Magnenat-Thalmann et al. (Eds.): New Adv. Virt. Hum.: Arti. Intel., SCI 140, pp. 85–112, 2008.
springerlink.com

3.2 System Context: The VirtualHuman System

The research reported here has been conducted as part of the VirtualHuman project (see `http://www.virtual-human.org/`). Our current demonstration system uses multiple virtual characters to enact a story-line for two human users, who can interact with up to three virtual characters. In the following, we will first present the architecture of the VirtualHuman system. Then we will introduce the scenario of the VirtualHuman system and discuss the main tasks for our context model.

3.2.1 The Architecture of the VirtualHuman System

The VirtualHuman demonstration system consists of three standard PCs (one for speech recognition[1] (ASR), one for 3D-rendering (graphic output), and one for the dialogue) and a 3D presentation system (consisting of two high-resolution video projectors). The two human users stand in front of two columns, each equipped with a microphone (Sennheiser) and a track-ball controlling a mouse pointer.

The basic architecture of the VirtualHuman system is partitioned into four main function blocks (see Fig. 3.1):

Input Devices—The input devices of VirtualHuman comprise two microphones (one for each human user) and two track-balls. The output of the two microphones is first routed to a standard audio mixer where the two signals are assigned to different audio channels (left and right). Then, the signals are routed to the Behringer Autocom device which is a 2-Channel Expander/Gate/Compressor/Peak Limiter with Integrated Dynamic Enhancer, De-Esser and Low Contour Filter. This device is usually employed for high-quality audio recordings in studios or on stage. However, we have found that this device can also help conducting open-microphone speech recognition in noisy surroundings. It can be used to suppress all background noise below a certain threshold, and to limit peaks in the audio signal without corrupting the signal. Especially the suppressed background noise drastically improves the performance of the ASRs.

NarrationEngine—The NarrationEngine (NE) of VirtualHuman monitors and controls the development of the interaction between users and characters. To this end, it can prescribe *goals* for the individual virtual characters. These goals trigger particular behavior and action sequences of the characters. Consider, for example, a goal like *greet* which causes a character to greet the character that is stated in the body of the goal. The NE also receives feedback from the individual characters whether they were able to fulfill the prescribed goals or not. The basic approach of the NarrationEngine is described in [Göbel et al., 2006].

CDE Controller—The dialogue and behavior controller of VirtualHuman comprises a set of *Conversational Dialogue Engines* (CDEs; see

[1] Since there are two users, two ASR engines run in parallel.

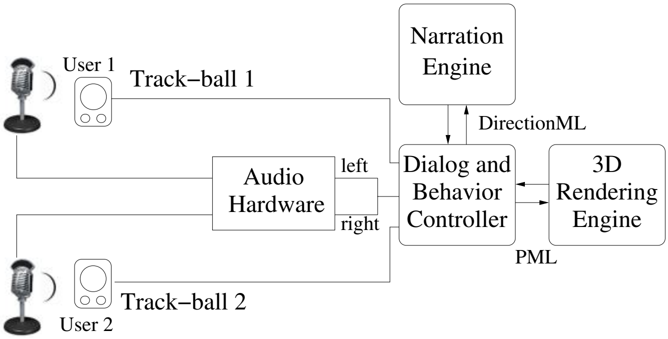


Fig. 3.1. Basic architecture of the VirtualHuman system

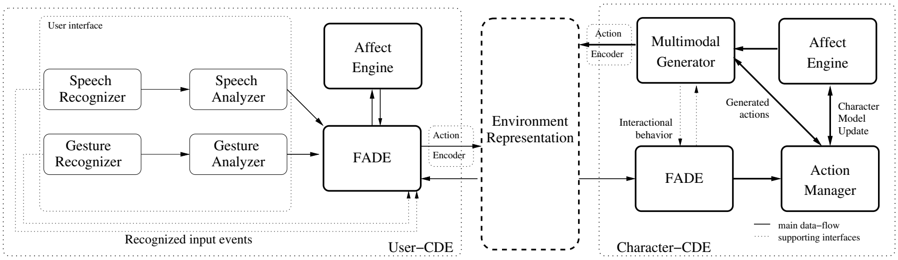


Fig. 3.2. Architecture of the User-CDE and Character-CDEs of the VirtualHuman system

[Löckelt and Pfleger, 2006]) which represent the individual participants of the scene. The dialogue controller takes the goals sent by the NE and distributes them to the individual CDEs (the language used for this information exchange is called *DirectionML*). We will discuss the CDEs in detail later in this section.

3D Rendering Engine—The 3D rendering engine of VirtualHuman is called *Avalon*[2]. Avalon is a high-performance 3D rendering engine that supports a variety of displays ranging from simple Web browsers to complex projection grids. The dialogue engine communicates with the Avalon player via PML (*Player Markup Language*) scripts which encode precisely timed sequences of verbal and nonverbal actions.

The CDE Framework. On the top-level of the CDE framework there is the *CDE controller* which controls and monitors the individual actions of the CDEs that participate in the interaction. Each participant of the interaction (virtual or human) is represented by its own CDE. Consequently, there are two different types of CDEs (see Fig. 3.2): (i) User-CDEs and (ii) Character-CDEs.

[2] See `http://www.zgdv.de/avalon/`

User CDEs. A User-CDE represents a human user of the VirtualHuman system. The task of a User-CDE is to convert the recognized verbal and nonverbal actions by the user into instances of the ontology that can be processed by the other CDEs. To this end, a User-CDE consists of a set of components: (i) a speech recognizer, (ii) a gesture recognizer, (iii) a natural language understanding (NLU) component and (iv) an instance of the context model called FADE (*Fusion and Discourse Processing Engine*).

Character-CDEs. A Character-CDE represents an autonomous virtual character that is part of the virtual world. The architecture of a Character-CDE comprises four main components (see Fig. 3.2): (i) an instance of FADE which is responsible for multimodal fusion, context-based discourse processing and reactive behavior, (ii) an action manager that is responsible for the deliberative actions of the character, (iii) an affect engine which maintains the affective state of the virtual character (see [Gebhard, 2005]) and (iv) a multimodal generator that generates synchronized verbal and nonverbal output of the character (see [Kempe et al., 2005], [Kempe, 2005]).

3.2.2 The Scenario of VirtualHuman

The basic scenario of the final VirtualHuman demonstrator is a football quiz show where two human contestants interact with three virtual characters: A moderator and two virtual experts named Ms. Herzog and Mr. Kaiser. The quiz consists of two phases; in the first phase the moderator shows short videos of suspenseful situations of past football games.

These videos stop just when the situation is about to be resolved (e. g., the striker attempts to shoot) and the human contestants have to guess how the situation will go on. In order to ease the task for the human contestants, the moderator provides three possible answers and the human contestants also have the opportunity to ask the virtual experts for their opinion. After three rounds of videos, the moderator selects the winner of this first phase (the left part of Fig. 3.3 depicts a screen-shot of the first phase of the quiz).

In the second phase, the winner of the first phase will be given the opportunity to assemble the line-up of the German national football team. The scenery of the game changes to the one shown in the right part of Fig. 3.3. The moderator and the female expert are positioned behind a 3D representation of a football field with marked positions for the players while the available football players are displayed in a list on the right side of the screen. The human user can place players either by means of spoken or multimodal commands (see the sample dialogue in the next section).

The game ends either after a predetermined time interval or when the user has assembled a complete football team. In any case, the moderator and the virtual expert will evaluate the assembled team and discuss the positive and negative aspects of the team.

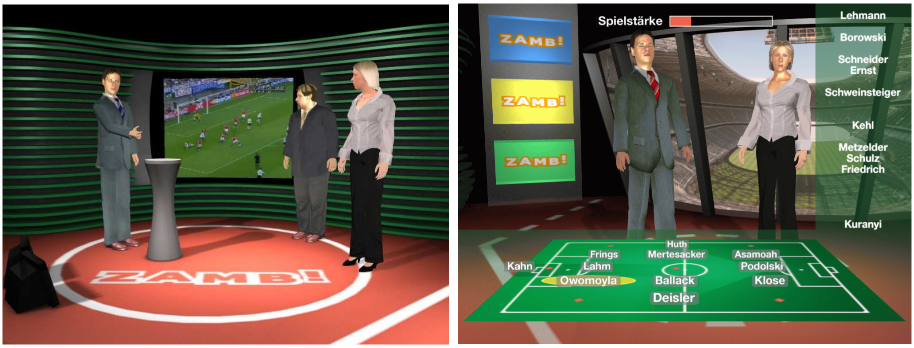


Fig. 3.3. The left part of the figure shows the second version of the VirtualHuman system during the first game phase. The right part shows game phase two.

An Example Interaction. Since a typical interaction with the VirtualHuman system lasts about 15 minutes, we will give here only two short fragments of the two game stages. For the first stage, we will start just when the moderator shows the first video sequence[3]:

> **Moderator:** [gazes at User 1 and User 2] *"Ok, here comes the first video."*
> [The video is shown on the projection screen behind the virtual characters]
> **Moderator:** [gazes at User 1 and User 2] *"What happens next: first [counting gesture], the goalkeeper will save the ball; second [counting gesture], the striker will score; third [counting gesture], the striker will miss the goal."*
> **Moderator:** [gazes at Ms. Herzog] *"Ms. Herzog, what do you think?"*
> **Herzog:** [gazes at the moderator] *"Well, I think the goalkeeper saves the ball."*
> **Moderator:** [gazes at User 1] *"Ok, User 1, what do you think?"*
> **User 1:** *"Hhm, I think Ms. Herzog is right."*
> **Moderator:** [gazes at User 1 and then at User 2] *"Well, I don't know, User 2, what do you think?"*
> **User 2:** *"I guess the striker scores."*

Then the moderator shows the complete video sequence and awards the points (a correct answer is worth one point). After three rounds of videos, the winner of the first stage (i. e., the human player with the highest score) will proceed to the second stage.

After a brief introduction to the line-up game, the moderator starts the second stage:

> **Moderator:** [gazes at User 1] *"Ok, let's get started."*
> **User 1:** *"Put Oliver Kahn into the goal."*
> **Herzog:** [nods; gazes at User 1] *"That's an excellent move!"*

[3] Note that the contributions of the virtual characters and the users were translated from the German original.

Moderator: [gazes at Ms. Herzog; nods; gazes at User 1] *"Great, Kahn in the goal position."*
User 1: *"Ms. Herzog, give me a hint!"*
Herzog: [smiles; gazes at User 1] *"I would definitely put Ballack into the central midfield."*
User 1: *"Ok, let's do that."*
Herzog: [smiles; nods; gazes at User 1] *"You won't regret this move."*
Moderator: [nods] *"Great, Ballack as central midfielder."*
User 1: ... [hesitates]
Moderator: [gazes at User 1; encouraging gesture] *"Don't be shy!"*
User 1: *"Hhm, put Metzelder to the left of Ballack."*

This interaction goes on until the user has assembled a complete team and indicates that he is finished, or until the user runs out of time. At the end of the game the moderator and the virtual expert evaluate the team and discuss potential problems for the individual parts of the team.

3.2.3 Context-Dependent Phenomena in VirtualHuman

The sample interactions show several challenging discourse phenomena whose interpretation requires for contextual knowledge. The following enumeration provides a brief overview of the key discourse phenomena the context model and FADE have been designed for:

Reference resolution

Resolution of spatial references

(1) **User:** *"Ballack right of Frings."*

(2) **User:** *"Ballack in front of Lehmann."*

For the resolution of spatial references, the context model needs to maintain a representation of the spatial organization of the football field displayed in the virtual studio. The actual reference resolution is then realized by FADE's built-in methods for reference resolution.

Resolution of discourse references

(3) **Herzog:** *"I think Ballack shoots the ball into the evening sky."*
Kaiser: *"I am sure Ballack scores."*
Moderator: *"Player 1, what's your opinion?"*
User: *"I think Ms. Herzog is right."*

The recognition result of the user's last utterance is the dialogue act *Agree* comprising the information that the user agrees with the character Ms. Herzog. However, this dialogue act still lacks the semantic content of what Ms. Herzog actually said and thus the task of FADE—that represents the moderator—is to resolve what has actually been said by Ms. Herzog before the action manager is able to process the *Agree*.

(4) **Herzog:** *"I would place Ballack into the central midfield."*
User: *"Ok, let's do that."*

This example basically requires the same functionality of FADE. The user reacts to the *Propose* dialogue act of Ms. Herzog with a *Confirm* which again needs to be enriched with the semantic content of Ms. Herzog's original utterance. This is realized via FADE's built-in methods for resolving discourse deixis.

Resolution of deictic references

(5) **User:** *"This player* [pointing gesture] *into the midfield."*

In this example, the user selects the player by using a deictic reference and an accompanying pointing gesture—a classic fusion task. Again, this functionality is realized by using FADE's built-in methods for reference resolution.

Addressee identification

(6) **User:** *"Ms. Herzog, what do you think?"*

In this case, the identification of the intended addressee is rather easy since the user has already provided the addressee through the vocative *Ms. Herzog*. However, the next example is more complex:

(7) **Moderator:** *"Tell me when you are ready."*
User: *"I am ready."*

In this example, the intended addressee can be inferred through the adjacency pair that is formed by the *Request* and *Confirm* dialog act. Thus, the addressee would be the previous speaker. However, if no information is present to narrow down the intended addressee, each character assumes they have been addressed. Then the characters have to decide based on their capabilities and current aims whether they will react or not. Consider, for example, the following utterance by the user in the second game stage:

(8) **User:** *"Put Michael Ballack into midfield."*

From the perspective of FADE it is not clear in this case whether the moderator or Ms. Herzog has been addressed. Consequently, the individual FADE components of the two characters assume to be the ones addressed. However, only the moderator's action manager (see below) is able to deal with the user utterance, and Ms. Herzog's action manager would ignore it.

Triggering backchannel feedback

Another important task for FADE in the VirtualHuman system is to trigger the gazing behavior of the individual characters. If, for example, someone starts speaking, it is natural for the listeners to react by gazing at the speaker in order to indicate their attention and their general acceptance of this participant as the speaker. Moreover, it is also important for the speaker to look at the intended addressee(s) at the end of the turn in order to encourage them to take over the turn.

3.3 A Comprehensive Context Model for Virtual Characters

This section presents and discusses a comprehensive context model for multimodal - multiparty dialogues. This model is designed to support both the task of multimodal fusion and the task of discourse processing in a multimodal dialogue system. We will start by giving an overview of the basic contextual categories of the model and then discuss the difference between the immediate conversational context and the discourse context. Section 3.3.2 and section 3.3.3 then discuss the structure of the conversational and the discourse context respectively, in detail.

3.3.1 Modeling Context

Based on the notion of *local context factors* discussed in [Bunt, 2000], we differentiate five categories of contextual factors that need to be represented in order to be able to deal with the full range of context-dependent multimodal discourse phenomena[4]:

Physical Context—objects that are present in the surroundings and that might serve as potential referents
Perceptual Context—general events or actions conducted by the other participants; projected/expected actions of the participants
Conversational Context—current status of the conversation with respect to turn management; current conversational roles of the participants (speaker, addressee, overhearer, eavesdropper)
Linguistic Context—discourse history; including unique representations of referents
Social Context—social roles of the participants.

Taking the Perspective of the Participants. Typically, the context model of a dialogue system represents context from a bird-eye perspective, i. e., just like an impartial recorder who tracks the individual actions of the participants. The key to our approach is that the context model is tailored to the view that each individual participant has of the interaction. This means that the context model will always reflect an incomplete representation of what happened, or as [Bunt, 2000] highlighted: “There is no room here for an ‘objective’ notion of context, since the participants’ communicative behavior depends solely on how they view the situation, not on what the situation really is” [Bunt, 2000, p. 101]. However, this subjective notion of context is only applicable for the discourse context. As we discuss below, there are aspects of context that provide an objective view (i. e., the representation of the physical surroundings).

[4] Note that even though the names of some categories resemble the ones used in [Bunt, 1994, Bunt, 2000], our connotation is slightly different.

Immediate Conversational Context vs. Discourse Context. The distinction between *interactional* and *propositional* information in contributions (see [Cassell et al., 1999]) is of particular importance for our context model as these two types of contributions require different processing strategies. Interactional information contributes to the structural organization of the conversation as it regulates the exchange of turns, helps to avoid overlapping speech, is used to provide backchannel feedback and supports the identification of the intended addressees of a contribution.

Based on this distinction, we differentiate between two types of context representations that an artificial participant (i. e., a dialogue system or a virtual character) in a conversation needs to maintain. The first is the *immediate conversational context* representing the current physical and perceptual context. This immediate turn context serves as a temporal storage for perceived monomodal events that need to be interpreted in their context of use (this approach is derived from [Pfleger, 2004]). The second type is a long-term *discourse context* representing previous contributions of the participants. This discourse model supports the resolution of referring expressions by means of referents derived from accompanying gestures and those introduced in the previous discourse.

While interactional nonverbal behavior (such as head nods, gazing, beat gestures) is incorporated into the representation of the immediate turn context, pointing gestures and iconic gestures are incorporated into the discourse context. These latter gestures are typically resolved by a multimodal fusion component but in our approach they are processed together with the spoken referring expressions. Formally, we define the context Γ of interactions as pairs of $< IC_A, DC_A >$, where IC_A is the immediate conversational context as it is perceived by a participant A, and DC_A is the discourse context as it is perceived by participant A.

3.3.2 The Immediate Conversational Context

The purpose of the immediate conversational context is to maintain an effective representation of the conversational status so that perceived monomodal events can be interpreted with respect to their impact on the interactional development of the conversation. The immediate conversational context comprises aspects of the physical context, perceptual context, conversational context and social context from the perspective of an individual participant of an interaction. The structure of the conversational context is centered around the physical context, the individual participants and their current actions. But this context represents more information than the simple presence of physical objects and participants in the surrounding. For example, besides the physical properties, the representations of the participants comprises also their current conversational role, their current active interactional signals (e. g., gaze, gesticulation) and their social status (if available). Additionally, the participants have a rich representation of themselves including their individual aims.

This conversational context also builds the basis for any reactive behavior of the system. All perceived monomodal and so far uninterpreted events are categorized and integrated into this context model. This permits direct reactions to events that are of particular impact for the participant that a context representations stands for. If, for example, someone else starts to speak, this is immediately registered in the conversational context which in turn prevents our participant from unintentionally interrupting the current speaker.

The purpose of the immediate conversational context is to always reflect the current state of the conversation. This has the consequence that it has to be updated as frequently as possible. Moreover, the conversational context does not reflect the history of events but rather provides a snapshot of the current state of affairs. Any information that is needed later on has to be stored in the discourse model.

Formally, we define the immediate conversational context IC_A of a participant A as a set of $< CS, P_i, ..., P_n >$, where CS represents the conversational status and P_i a participant of the conversation. In the remainder of this section, we will describe in detail the model of the immediate conversational context. However, note that this information is to some extent tailored to the needs of the task at hand and is subject to extensions or reductions if other tasks require more or less details.

Representing Dialogue Participants. Each participant of an interaction typically has a good notion of the other participants present and uses this information in various ways both for processing and generating contributions. Part of this information can be directly perceived from the participants' appearance (e. g., the sex, sometimes the social status) while other information can only be projected or collected during the course of the interaction. But more importantly, only parts of this information is really relevant for conducting the interaction itself.

The fact, for example, that someone is left-handed or has black hair usually has no direct impact on the conversation. But this leaves us with the question of what is relevant information about the participants of a dialogue? The answer to that question, however, is application specific and thus we have to consider the phenomena and tasks at hand in order to determine the types of information that have to be stored for each participant.

When we start with the addressee identification, the name and sex of a participant support the identification of the current roles of the participants so that this information should be part of the model. But of course, gazing behavior also contributes a lot to determining the intended addressee of an utterance. In general, any nonverbal behavior of a participant should be represented in the conversational context, as these are considered strong turn-taking signals. Regarding place deixis or spatial references, the participant's preferred or currently employed frame of reference is an important information as speakers tend to tailor their contributions to that of the addressees and use the same frame of reference. Additionally, the current location of a participant influences the resolution of spatial referring expressions and should therefore be represented in the conversational context.

Membership in communal groups (see, e. .e, [Clark, 1996]), expertise and the social status of a participant also contribute to the processing and generation of dialogue contributions but in a less direct way than the aforementioned aspects. This information is used to select appropriate and adapted referring expressions and phrases. Moreover, the perceived emotional state of a participant also has some impact on the interpretation of an utterance (i. e., irony, sarcasm).

To summarize, the representation of a participant in the immediate conversational context comprises the following aspects:

Name: The participant's first and last name (but also nicknames).
Sex: The participant's sex (*male*, *female*).
Nonverbal behavior: The currently active nonverbal behavior represents the list of currently perceived nonverbal behavior of that participant.
Frame of Reference: The currently used frame of reference: intrinsic, relative or absolute (see [Levinson, 1983]).
Position: The participant's position in the scene (i. e., the top-level physical environment).
Emotional state: The participant's emotional state (if available).
Communal groups: Assumed membership in communal groups.
Expertise: Assumed expertise of the participant.
Social status: Assumed social status of the participant.

If the participant to be represented is a human user, part of this information is represented in so-called *user models* (see for example [Wahlster and Kobsa, 1989], [Heckmann, 2006]). Usually, a user model encompasses information and assumptions about all aspects of the user that might affect the interaction. Many dialogue systems comprise an explicit *user modeling* component that provides this kind of information for other components. In that way, the representation of a human participant in the immediate conversational context can be provided by a user model.

Modeling the Conversational Status. While interacting with other people, humans usually have a clear understanding of the conversational status, i. e., they know who the speaker is, what their own role is, or if they are supposed to speak. This understanding is crucial for a successful interaction between two or more interlocutors. To this end, our context model comprises a representation of the *conversational state* which can be filled with a wide variety of information.

Participating in multiparty interactions requires a participant to carefully monitor the turn-taking signals of the current speaker (e. g., pauses, rising or falling pitch). Moreover, in order to be able to identify the addressee(s) of an utterance, a participant also needs information about the current and previous speakers and addressees. In order to support the corresponding reasoning processes, we envision the following information in the conversational status of our context model:

Characteristics of the current turn:
Turn duration—how long is the current speaker already holding the turn.

FrameOfReference—holds currently active frame of reference.

TemporalReferencePoint—currently activated temporal reference point; the last time point mentioned in the interaction.

Prosody—represents falling pitch at the end of a sentence or the drawl of a syllable at the end of sentence.

Current speaker—describing the current speaker (not set if the floor is available, i. e., nobody is claiming the speaking turn).

Current addressees—describing the current addressees as far as they can be inferred.

Current overhearers—describing the participants that are classified as overhearers.

Current bystanders—describing the participants that are classified as bystanders.

Previous speaker—holding the participant that was the speaker in the previous turn.

Previous addressees—holding the participants that were addressed in the previous turn.

The kind of information represented in the conversational context also depends on availability issues (e. g., in some cases the sensoric devices are not yet able to capture the information) and for some system configurations it simply does not make sense to model the information (e. g., in a classic dialogue system there are only two participants, the user and the system, which makes the representation of bystanders and over-hearers obsolete). Thus, the conversational status is not a fixed model but rather flexible with respect to the needs and capabilities of the dialogue system.

Modeling the Physical Environment. A key aspect of situated multimodal communication is the need for a comprehensive representation of the physical context within which the interaction takes place since people integrate both verbal and nonverbal references to the physical environment into their contributions (see, e. g., [Knapp and Hall, 2002]). Consequently, a comprehensive context model needs to incorporate a detailed model of the physical environment, including all objects, persons and other entities that are in the visual field of the dialogue participants and that might be referenced during the discourse. It is also important to model the relations between the Physical Objects by means of differentiation criteria like size, color and shape (see, e. g., [Salmon-Alt, 2000]). These differentiation criteria can be used to non-ambiguously identify a referent out of a number of similar referents.

Another key aspect of this model of a physical environment is that it supports the cascading of embedded physical environments. This feature enables the modeling of distinct areas that themselves represent a closed physical environment (e. g., lists that are displayed on a screen, or rooms that are part of a physical environment modeling a house or an apartment). Once an object of the physical environment has been referenced, the object will also appear in the discourse context (see below). The concept of physical environments we have

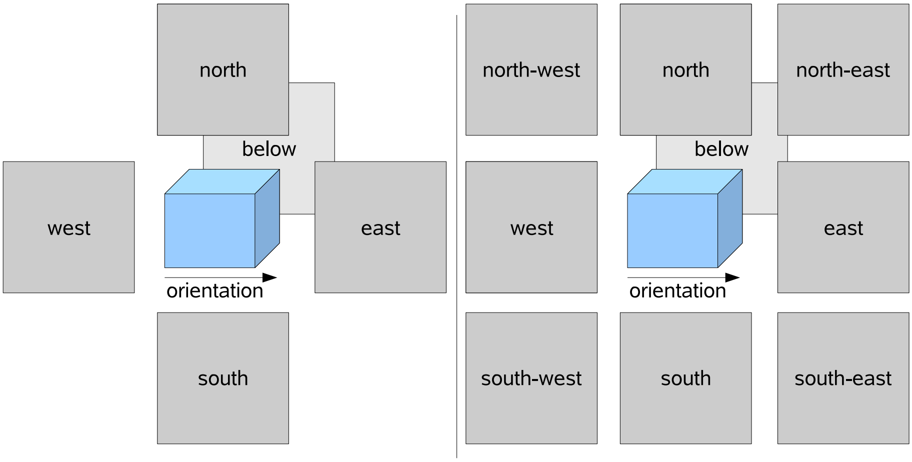


Fig. 3.4. Graphical representation of the spatial relations between an object of the physical environment and its neighboring objects. The left part of the figure shows the coarse-grained model using six relations and the right part shows the fine-grained model using ten relations. Due to layout restrictions, both figures do not show the *above* relation.

developed for this work is comparable to the concept of *domains of references* of [Salmon-Alt, 2000]. However, it is more flexible since it supports a spatial representation of the objects located in the physical surroundings.

Representing Basic Spatial Relations. Representing spatial relations between objects in the physical environment has a long research tradition and a number of comprehensive approaches and models have been put forward (see for example [Maass, 1996], [Gapp, 1996], [Blocher, 1999], [Fuhr et al., 1995]). In what follows, we will describe a model for representing spatial relations between Physical Objects (see Sec. 3.3.3) that is tailored to the task of resolving spatial references for multimodal dialogue systems. This model is based on labeled spatial relations between objects but does not incorporate as many details as other approaches.

We model objects that are located within a closed physical environment by means of either six or ten spatial relations, *northernNeighbor, north-easternNeighbor, easternNeighbor, southernNeighbor, south-easternNeighbor, westernNeighbor, north-westernNeighbor, above* and *below* plus the current orientation of each object (*north, east, south* and *west*). Fig. 3.4 gives a graphical representation of these two types of spatial models. The concept of labeled spatial relations resemble the *spatial prepositions* used in [Fuhr et al., 1995].

Our model is organized as follows: each object located in the scene is represented by means of an *AbsolutePosition*. An AbsolutePosition is represented in our meta-ontology but for now it is sufficient to think of a typed feature structure that comprises a set of features. The most important features of this structure are:

- Feature *ontologicalInstance*—this feature contains the ontological instance representing the object that is described by an AbsolutePosition.
- Feature *coordinates*—this is an optional feature: If the perception components are able to provide two-dimensional or three-dimensional coordinates describing the location of the object with respect to a fixed coordinate system, this slot will contain that information.
- Feature *orientation*—this feature describes the current orientation of the object; valid values are: *north, east, south, west.*
- Feature *northernNeighbor*—this feature contains a link to the AbsolutePosition of its northern neighbors (can be empty if there is no northern neighbor).
- Feature *easternNeighbor*—this feature contains a link to the AbsolutePosition of its eastern neighbors (can be empty if there is no eastern neighbor).
- Feature *southernNeighbor*—this feature contains a link to the AbsolutePosition of its southern neighbors (can be empty if there is no southern neighbor).
- Feature *westernNeighbor*—this feature contains a link to the AbsolutePosition of its western neighbors (can be empty if there is no western neighbor).
- Feature *above*—this feature contains a link to the AbsolutePosition of its neighbors above (can be empty if there is no neighbor above the object).
- Feature *below*—this feature contains a link to the AbsolutePosition of its neighbors below (can be empty if there is no neighbor below the object).

This means that each *AbsolutePosition* that represents an object in the scene also represents the spatial relations the object currently fulfills with respect to the organization of the scene from an absolute point of view (using viewpoint neutral descriptions). If we consider, for example, a very simple physical environment consisting of a football team lineup, the idea of this representation will become clearer. Figure 3.5 depicts the spatial arrangement of the individual football players. Given this scene the representation of the physical environment looks like this:

$$\begin{bmatrix} \text{PhysicalEnvironment} & \\ \text{absolutePosition:} & \boxed{1} \begin{bmatrix} \text{AbsolutePosition} & \\ \text{ontologicalInstance:} & \begin{bmatrix} \text{FootballPlayer} & \\ \textit{name} & \text{“}\textit{Ronaldo}'' \\ \ldots & \end{bmatrix} \\ \text{easternNeighbor:} & \boxed{2} \\ \text{northernNeighbor:} & \boxed{3} \end{bmatrix} \\ \text{absolutePosition:} & \boxed{2} \begin{bmatrix} \text{AbsolutePosition} & \\ \text{ontologicalInstance:} & \begin{bmatrix} \text{FootballPlayer} & \\ \textit{name}: & \text{“}\textit{RobertoCarlos}'' \\ \ldots & \end{bmatrix} \\ \text{westernNeighbor:} & \boxed{1} \\ \text{easternNeighbor:} & \boxed{4} \\ \text{northernNeighbor:} & \boxed{5} \end{bmatrix} \end{bmatrix}$$

Types of Spatial Organizations. Typically, the individual spatial organizations represent logical groups of objects that belong together either because of their

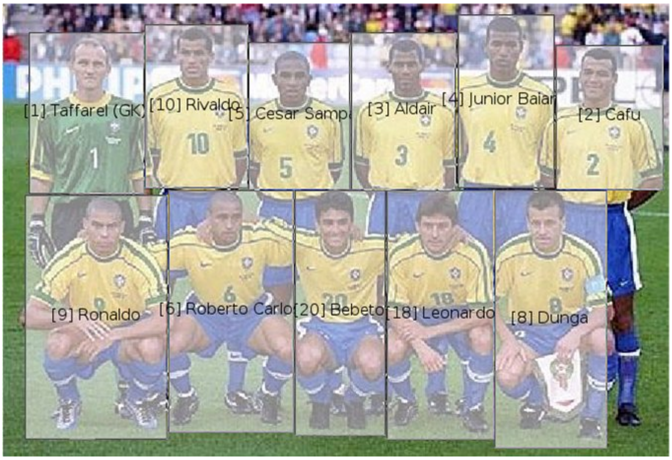


Fig. 3.5. Example configuration of a physical environment: The football player Ronaldo is either to the right of Roberto Carlos or to the left of him, depending on the point of view

spatial assembly or because of shared features like type, color and size. Examples of such groupings are the players on a football field or the elements of a list. There is virtually an unlimited number of possible organizations such as lists, double-column lists, set football fields, group photos and maps. Each type of spatial organization comes with its own vocabulary that is typically used for place deixis or spatial references. The entries of a single-column list, for instance, can be referenced to by expressions like *the first*, *the second* or *the last*. However, some spatial organizations require more complex expressions, as for example references to a double-column list also require the determination of the row, e. g., *the first one on the right*. Other examples of complex spatial organizations are a football field or a map.

3.3.3 The Discourse Model

The purpose of the discourse model is to gain a comprehensive representation of the contributions to the propositional layer of an ongoing interaction to enable the resolution of referring and elliptical expressions. Given all the possibilities of referring expressions as they occur in natural interactions, our goal is to design a context model that addresses the two central challenges of reference resolution:

1 What information should the system maintain that might be useful for resolving future references?
2 Given an anaphoric reference in a follow-up sentence, how can we pick out the entity it is meant to represent?

The basis of our discourse model is the model we developed for the SmartKom system (see [Pfleger, 2002] for an overview of that model). Following the *three-tiered discourse representation* of [LuperFoy, 1991], our discourse model comprises two main layers: (i) a Modality Layer—extending her linguistic layer and (ii) a Referential Layer—extending her discourse layer. The third layer—the knowledge base or belief system—corresponds in our approach to the long-term memory. Besides this, the discourse context also comprises a sequential representation of the course of the interaction, i. e., a discourse history. This discourse history encompasses information about the speaker that performed the utterance and the propositional content. Formally, we define the discourse context DC_A of a participant A as a tuple of $< ML, RL, LTM, DH >$, where ML corresponds to the Modality Layer, RL corresponds to the Referential Layer, LTM corresponds to the long-term memory, and DH to the discourse history.

In the remainder of this section, the first two layers of this discourse model will be discussed in detail. The long-term memory will be introduced in the next section.

Modality Layer. The objects of the Modality Layer provide information about the surface realizations of objects at the Referential Layer that have been introduced into the discourse. Thus, the Modality Layer describes the circumstances that caused the increase in activation of their corresponding instance. The Modality Layer consists of three classes of objects reflecting the modality by which the corresponding Referential Object was referenced: (*i*) *Linguistic Actions*, (*ii*) *Nonverbal Actions*, and (*iii*) *Physical Objects.*

Linguistic Actions: Linguistic Actions resemble the linguistic objects of [LuperFoy, 1991]. They comprise information about the surface realization of an instance like lexical information (the lemma used to reference an instance), syntactical information (e. g., number, gender or case), its realization time or the type of reference (e. g., definite/indefinite, deictic/anaphoric/partial anaphoric). Each Linguistic Action is linked to exactly one instance of the knowledge base and when this link is established, the activation of the referenced object is increased. Linguistic Actions are of particular importance for the resolution of referring expressions as they provide the linguistic information needed to identify co-references on the linguistic level.

A Linguistic Action is described by the following features: (i) the point in time at which the referring expression was uttered, (ii) the lemma used to reference the corresponding instance, (iii) the syntactical information (i. e., number, gender and case of the referring expression), (iv) the type of the reference (i. e., definite, indefinite, deictic, anaphoric, or partial) and (v) the participant that realized the action.

Nonverbal Actions: Nonverbal Actions represent the nonverbal behavior of the interlocutors that contribute to the propositional content of the utterance (e. g., pointing gestures, iconic gestures, emblematic gestures, but also gaze behavior or drawings). Nonverbal Actions comprise information about the type of nonverbal action, its start and ending time. Nonverbal Actions facilitate the resolution of deictic expressions (e. g., *"What's the name of that [pointing gesture] player?"*).

A Nonverbal Action is described by the following features: (i) the point in time at which the nonverbal action was displayed, (ii) the duration during which the nonverbal action was displayed, (iii) an ontological representation of the nonverbal action (e. g., deictic, iconic, emblematic, gaze), (iv) a link to the referenced object at the Referential Layer and (v) the participant that realized the action.

Physical Actions: Physical Actions describe changes of the state of the physical world (e. g., the appearance or disappearance of objects in the physical environment or intentional actions) performed by a participant of the discourse. They comprise information about the type of the event, when it happened, and about the spatial properties of that object (including its relative position to other objects in the scene).

A Physical Action is described by the following features: (i) its realization time, i. e., the time point when the Physical Action began, (ii) the duration of the Physical Action, (iii) an ontological representation of the Physical Object at the Referential Layer that is involved in the Physical Action, (iv) an ontological representation of the type of the physical action (e. g., *Appear, Disappear*), (v) a representation of the location of that object represented by means of an absolute position and (vi) the participant that realized the action.

Referential Layer. Objects at the Referential Layer provide the link to the instances of the long-term memory (see section 3.3.4). Each object at the Referential Layer (if completely disambiguated) represents a unique instance of the long-term memory whose activation value exceeds the threshold that differentiates between activated and in-activated objects (see section 3.3.4). We distinguish three types of objects at the Referential Layer: (i) Discourse Objects, (ii) Implicitly Activated Objects and (iii) Physical Objects.

Discourse Objects: Discourse Objects are containers for instances that were directly mentioned during the preceding discourse. They comprise a unified representation of the semantic information gathered so far. In case a Discourse Object is completely resolved, its unified representation is replaced by a link to the corresponding instance of the long-term memory. Additionally, it contains a set of links to instances of the long-term memory (more than one in case of ambiguous/under-specified references) and links to objects at the Modality Layer; every time a Discourse Object is mentioned, a new link is added.

Implicitly Activated Objects: Implicitly Activated Objects are objects that are activated by means of the mentioning of a Discourse Object that is associated with them (e. g., instances like the pilot and the passengers in the context of the mentioning of a plane). If a Discourse Object accesses an instance in the LTM, the activation of instances related to it is increased by a dynamic factor which depends on the activation of the superordinated instance and the *strength* of the relation between them. The spreading of activation is a recursive process (see section 3.3.4). Implicitly Activated Objects may appear in the discourse context when their corresponding Discourse Object appears. This happens in case their activation exceeds the threshold. Also, the activation of Implicitly Activated Objects decreases faster than that of Discourse Objects. Consequently, they are only accessible for a short time.

Physical Objects: Physical Objects represent objects that can be perceived from the visual environment. If a Physical Object is explicitly activated through the mentioning of a Discourse Object, it can serve as a referent for a referring expression. Physical Objects are not only part of the Referential Layer as they are also part of a superordinate structure representing the complete physical surroundings by modeling the relations between the Physical Objects located in a scene (e. g., the grey building is to the left of the blue building).

3.3.4 The Long-Term Memory

A key feature of our context model is the existence of a long-term memory (LTM) that is connected with the working memory (in psychology it is currently also common to view the working memory as a distinct part of the LTM). This memory unit provides access to all available but not directly accessible knowledge of the dialogue system. The objects represented in our LTM form a semantic network where the individual instances are connected through relations (i. e., the properties of the ontological classes). The left part of figure 3.6 depicts a small excerpt of such a semantic network. At the bottom of this figure there is a instance representing the German football player *Michael Ballack*. This instance exhibits several connections with other instances like teammates, or the German national team that participated in the game against the USA at the World Cup in 2002.

An important aspect of this LTM is that every object has an activation value defining its accessibility. The higher the activation value, the easier it is to access the object (i. e., to retrieve the object from the complete set of knowledge). To account for the activation of neighboring instances that can be observed in human interactions, the activation of a knowledge chunk is passed on to its associated chunks by a process called *spreading activation*. Spreading activation means not only that each connected object receives part of the activation of its neighbors, but also that it spreads its own activation on to its neighbors. The task of the LTM within this context model is to provide access to the actual referents for referring expressions including their associated entities. This information can be used for the resolution of implicit referring expressions.

All entities of the long-term memory are at first not accessible, i. e., they are not directly accessible by means of referring expressions. The status of an inactivated entity only changes if its activation exceeds a certain threshold that marks the border between accessible and inaccessible entities. All entities whose activation is below this activation threshold are not accessible for the reference resolution process. However, if instances receive additional activation by means of the spreading activation process, they can exceed the activation threshold which makes them accessible for the reference resolution process.

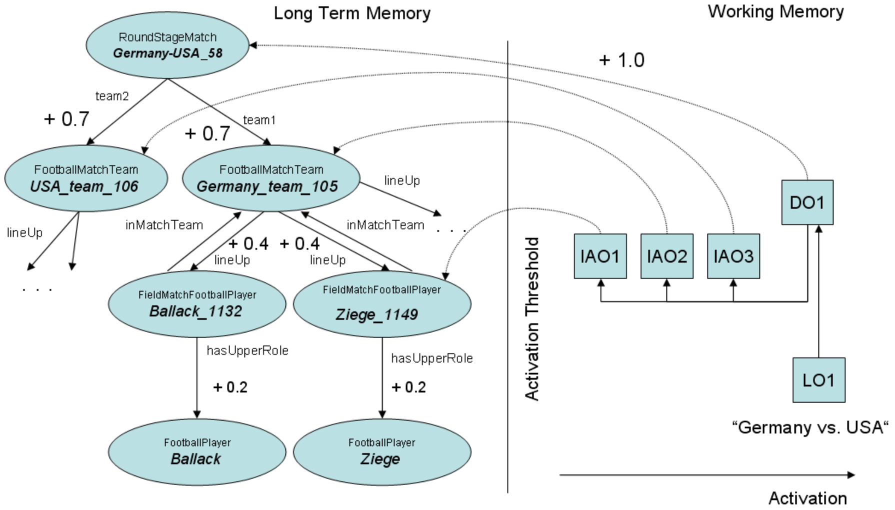


Fig. 3.6. Basic organization of the context model (taken from [Pfleger and Alexandersson, 2006:]) The left part of the figure shows an excerpt of the long-term memory and the right part shows some objects on the Referential Layer in the working memory. Both memory structures are separated by an activation threshold.

3.3.5 The Context Changing Function of Utterances, Physical Actions and Events

As discussed at the beginning of this section, every communicative or physical action that is uttered within the course of an interaction has to be interpreted in the light of its context but also changes this context. Formally, this is realized by means of a context update function *F(p,i,s,a)* that takes the propositional content p, the interactional content i, the speaker s, and addresses a of an utterance or a physical action and updates a given context Γ to Γ’. In the following two subsections, we will discuss how this is achieved for the immediate conversational context and discourse context, respectively.

Besides utterances and physical actions, we distinguish a tuple of four basic events that also have a context changing function: (i) *Join(a)*—a participant a joins the conversation, (ii) *Leave(a)*—a participant a leaves the conversation, (iii) *StartOfTurn*—a participant takes the turn and (iv) *EndOfTurn*—a participants ends a turn. These basic events are also incorporated into the context by means of four context update functions $Join$, $Leave$, $StartOfTurn$ and $EndOfTurn$ that all take an agent a as argument in order to update a given context Γ to Γ’.

Updating the Immediate Conversational Context. As discussed in section 3.3.2, the conversational context IC_A of a participant A is represented by the set $< CS, P_i, ..., P_n >$, where CS is the conversational status and $P_i, ..., P_n$ are the registered participants. There are two primary update functions for dealing with changes in the number of present participants. The context update function $Join(a)$ updates the list of registered participants by adding the participant a. Accordingly, the function $Leave(a)$ removes the participant a from the list of registered participants. Additionally, the context update function $StartOfTurn(a)$ registers the specified agent a as the current speaker of the interaction while the function $EndOfTurn(a)$ changes a's status from current speaker to previous speaker.

When a participant A performs a nonverbal action NA, the immediate conversational context is changed so that the representation of that participant (e. g., P_A) will comprise the performed nonverbal action in the field *Nonverbal behavior* for the duration of the nonverbal action.

Updating the Discourse Context. The discourse context is defined as the set $< ML, RL, LTM, DH >$, where ML corresponds to the Modality Layer, RL corresponds to the Referential Layer, LTM corresponds to the long-term memory, and DH to the discourse history. When a participant P_B performs an utterance, all four aspects of the discourse context have to be updated. The discourse history is updated by adding the propositional content together with information about the speaker at the end of the sequence. The other three layers are extended as follows: First, the propositional content p has to be analyzed with respect to the individual Referential Objects that are referenced in p.[5] For each referenced Referential Object it is first determined whether it has already been mentioned during the previous discourse. If this is not the case, the Referential Layer will be extended with this Referential Object. Otherwise, the existing Referential Object will be used to link it with the corresponding Modality Object (that is either a Linguistic Object in case of a verbal reference, a Gestural Object in case of a gestural reference or a Physical Action in case of a physical action).

In a second step, the activations of the objects at the Referential Layer and in the long-term memory that have been mentioned during a turn are increased accordingly. In contrast, the activation of all Discourse Objects at the Referential Layer that have not been mentioned in the current utterance will be reduced and all objects whose activation is below 0 will be removed from the Referential Layer, which means that they are no longer accessible. Finally, the spreading activation process of the long-term memory ensures that the Implicitly Activated Objects at the Referential Layer will be updated.

3.4 Applying the Context Model

The context model discussed in the previous section has been implemented in the FADE component of the VirtualHuman system. FADE consists of two processing layers (see Fig. 3.7): (i) a production rule system (called PATE; *a Production rule system Based on Typed Feature Structures*; see [Pfleger, 2004],

[5] Note that a Referential Object can be referenced by means of both verbal and nonverbal actions.

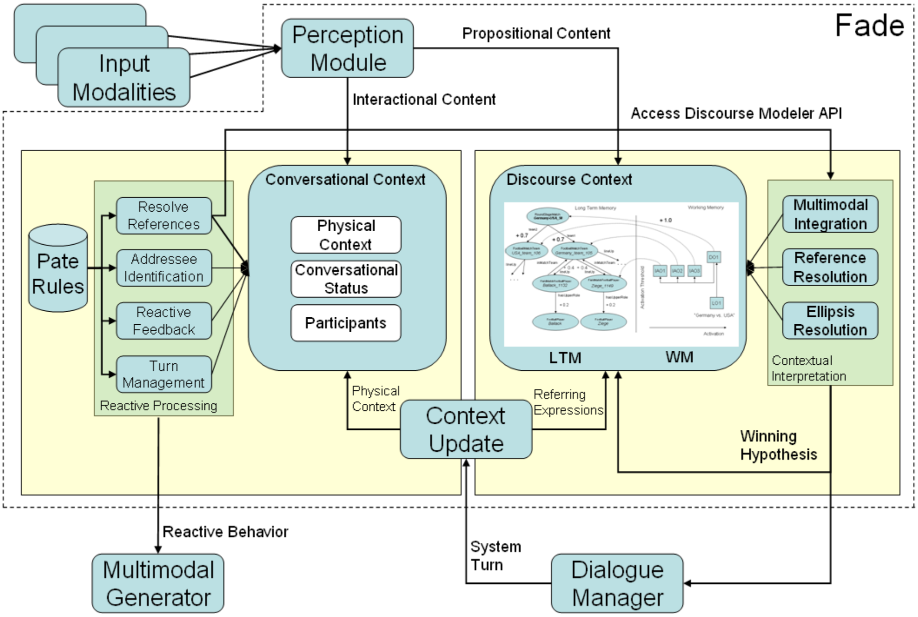


Fig. 3.7. The functional architecture of FADE. The left part of the figure shows the conversational context that is accessed and updated by PATE rules and the right part shows the discourse context and its API for processing propositional contributions.

[Pfleger and Schehl, 2006]) that is responsible for the reactive interpretation of perceived monomodal events, and (ii) a discourse modeler (called DiM) that is responsible for maintaining a coherent representation of the ongoing discourse and for the resolution of referring and elliptical expressions (see [Pfleger et al., 2003]).

3.4.1 The Basic Architecture of FADE

Making sense of perceived monomodal events consists of two aspects: (i) interpreting interactional signals in order to trigger appropriate reactions, and (ii) the integration of monomodal contributions that contribute to the propositional content of the turn. The Perception Module distinguishes the incoming monomodal events and updates the immediate turn context and the DiM. Key to our approach is that all processing instructions necessary to interpret the interactional events can be expressed through production rules. The remaining integration task is handled by the discourse modeling subcomponent.

In the following we will give a short overview of how the general processing strategies for the context dependent phenomena mentioned in Sec. 3.2.2 are realized.

3.4.2 Understanding and Generating Turn-Taking Signals

It is crucial for a participant of a multi-party interaction to understand turn-taking signals displayed by the other participants, as well as to display appropriate signals for the other participants. Moreover, timing has a great impact on the naturalness of this behavior. A backchannel feedback that is realized only a little too late might interrupt or at least confuse the current speaker and cause a temporary break-down of the turn-taking system.

For the current version of the VirtualHuman system we focused on the reactive generation of gaze behavior. A participant that perceives, for example, the onset of a verbal contribution of another character usually (but not always) reacts by gazing at the speaker. This is realized by means of a set of specialized production rules of FADE. If appropriate, FADE directly sends a request to the multimodal generator without consulting the action manager. The speaker also displays gaze behavior, however, with slightly different intentions. Speakers, in turn, gaze alternately at the participants who they want to address.

3.4.3 Resolving Discourse Deictic References

The resolution of discourse deictic references requires a comprehensive representation of the ongoing dialogue since not every reference refers to its immediate predecessor. The discourse deictic reference in turn (7) of the example dialogue in Sec. 3.2.2 is resolved by the characters as follows: First, they access their sequential dialogue history and try to retrieve the last contribution of the character Herzog. Then they need to integrate the proposal of Ms. Herzog with the actual utterance of the user.

3.4.4 Resolving Spatial References

Resolving spatial references depends on the point of view the speaker takes to encode the referring expression. This point of view is called the *frame of reference* [Levinson, 2003]. The frame of reference a speaker takes directly influences the selection of a particular referring expression, e. g., everything that is on my left is on the right of someone standing in front of me. [Levinson, 2003] distinguishes three main frames of reference: *intrinsic*, *relative* and *absolute*. When using an intrinsic frame of reference, the speaker takes the point of view of the relatum (i. e., the object that is used to locate the target object). In a relative frame of reference, the speaker takes an outside perspective (e. g., his own point of view, or that of someone else). Within an absolute frame of reference, everything is located with respect to the geographic north. While the latter frame of reference is always unambiguous the former two might introduce some ambiguities that need to be resolved.

The resolution of referring expressions involves the following aspects: (i) an up-to-date representation of the physical environment, (ii) knowledge of the currently active type of frame of reference and (iii) a mapping function that converts spatial references to locations or objects in the scene. In order to resolve spatial references, FADE first determines the currently activated physical environment and its corresponding active frame of reference and then maps the referring expression to an absolute location. If, for example, the user commands the system to *"Put Metzelder to Ballack's left"*, the system first searches for the current position of the player *Ballack* in the physical environment. Then it retrieves the orientation of that player and maps the referring expression to one of the absolute identifiers. At this point we assume a currently active frame of reference of type *intrinsic*, otherwise the system would need to determine the orientation of the speaker and then compute the mapping. In any case, the mapping function takes the referring expression (*left-of*) and the orientation of the relatum (*eastern*) which would result in an offset of 1. This means, we need to go *one* neighbor feature further to get the correct neighbor given the orientation. Normally (i. e., if the player would be oriented to the north), left-of would be mapped to the western neighbor, however, in our case we need to go one neighbor further which is the northern neighbor. If the player faces westwards, the mapping function would return an offset of 3 which means left-of is now the southern neighbor.

3.5 The Action Manager

In this section, we give a brief summary of how the action manager processes the output of FADE. After the input for a CDE has been fusioned and contextually disambiguated, it is passed on to the action manager module. This input always is in terms of *dialogue* or *physical acts* that are taken to occur in the context of a set of current activities performed by the characters.

The running activities are organized in leveled process hierarchy, as shown in Fig. 3.8. The top layer is concerned with the activities at different stages of the task, such as greeting the user, or playing the quiz. Activities may also comprise a number of sub-activities, e. g., the quiz consists of several consecutive turns. Activities interact with the environment (and the user) by executing *dialogue games* (constituting the second level) that are rule-based exchanges of dialogue acts and physical acts. A question-response game, for example, may consist of a *Question* dialogue act by an initiator followed by an *Response* by the addressee. Like activities, dialogue games can also involve sub-games; in the example, this could be a counter-question embedded in a question-response game. The bottom level in the figure illustrates that the actual surface realization of an act, as realized by the multimodal generator module or produced by the user, can involve several items in different modalities. In the latter case, the items are merged into one act by multimodal fusion before entering the action manager.

Dialogue games are instances of *game types* that are templates for classes of interactions of similar structure in terms of *moves* that represent one dialogue act exchange each. The possible move sequences in a game type can be depicted as a kind of finite state automaton, like shown in Fig. 3.9. A running activity creates a concrete instance of a game type by providing parameters that match the current situation, e. g., the actual content of a question and the addressee. The transitions between the states, as well as the whole of a game, are annotated

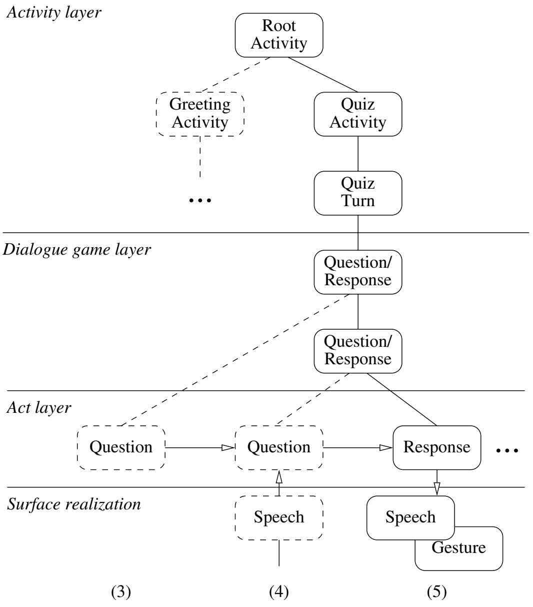


Fig. 3.8. Example process hierarchy

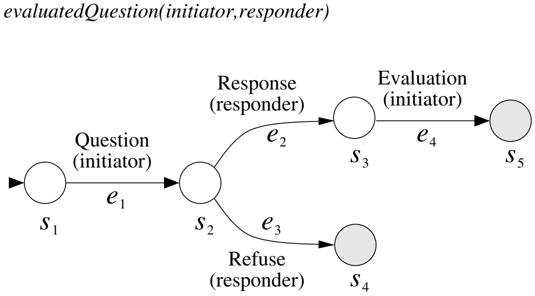


Fig. 3.9. Game type for an evaluated question-response

with preconditions and postconditions over the knowledge base of the character that allow to use a game type as an operator for a planning algorithm. Using the transitions with satisfied conditions, the action manager can also determine the set of expected response moves from other dialogue participants. FADE is notified of these expectations to help with disambiguating future input.

The set of processes that are executed by a CDE at any given time is determined on one hand by a Narration Engine module that is responsible for bringing the underlying story forward by setting goals for the characters; on the other hand, a user can also trigger new processes with communicative acts. If an input act by a user is recognized to fall outside a current activity, the action manager tries to start an activity that accommodates it. An example for this is when the user is expected to answer a quiz question, and instead asks one of the moderators for help. The narration engine is also notified of progress and successful or unsuccessful completion of goals.

The action manager's operation is explained in more detail in [Löckelt, 2005].

3.6 Conclusion and Future Work

In this chapter we have presented a comprehensive context model for multimodal multi-party discourse. At the top-level, this context model differentiates between an immediate conversational context and a discourse context. As discussed, it is important for the interpretation of interactional contributions to consider the immediate status of the conversation and the current conversational role of the participants which are provided by the immediate conversational context. The discourse context realizes a discourse history consisting of the individual propositional contributions of the participants. This discourse history centers around a representation of the individual referents of a discourse and comprises not only verbally mentioned instances, but also those that are introduced through nonverbal actions.

The second key feature of this context model is the associative long-term memory. A long-term memory is usually not considered to be part of a classic context model. However, as we have seen in the previous sections, some processes in the human LTM have a direct impact on the organization and structure of the contextual model. To this end, our approach integrates a structure resembling the human long-term memory. Moreover the activation spreading within the LTM ensures that not only the explicitly mentioned knowledge is accessible but also associated knowledge. We consider this to be the primary key for the resolution of implicit references.

Future Work. Even though conversations are organized in turns, this does not mean that only a single participant can speak at the same time. In fact, conversations are characterized by a great amount of overlapping speech without violating the turn-taking protocol. Mostly, this is feedback provided by the listeners/addressees to inform the speaker about their current understanding of the ongoing turn–this is called *backchannel feedback* [Yngve, 1970]. Backchannels can be expressed through both verbal (e. g., *"yes"*, *"ok"*, *"hmm"* etc.) and nonverbal behavior (e. g., head nods, facial expressions, etc.). As [Knapp and Hall, 2002] highlights, those responses can affect the type and amount of information given by the speaker. Another interesting observation is that speakers seem to implicitly request backchannel feedback from their audience as they organize their contributions in so called *installments*—each one separated by a short pause inviting the hearers to give some feedback [Clark and Brennan, 1991].

Currently, the characters display only gaze behavior while someone else is speaking, however, we want to add real backchannel feedback. We plan to add a small component that is able to identify short pauses in the speech signal of the user. The virtual characters in turn will use this information to generate appropriate backchannel feedback (e. g., slight head nods, facial expressions, etc.) depending on their current affective state and their understanding of the conversational state. Moreover, we also plan to extend the backchannel feedback of the characters when another character is speaking.

Acknowledgments

This research is funded by the German Ministry of Research and Technology (BMBF) under grant 01 IMB 01A. The responsibility lies with the authors. A brief version of this chapter was published in the IVA 2006 proceedings (see [Pfleger and Löckelt, 2006]).

References

[Blocher, 1999] Blocher, A.: Ressourcenadaptierende Raumbeschreibung: Ein beschränkt-optimaler Lokalisationsagent. PhD thesis, Department of Computer Science, Saarland University (1999)

[Bunt, 2000] Bunt, H.C.: Dialogue pragmatics and context specification. In: Bunt, H., Black, W. (eds.) Abduction, Belief and Context in Dialogue. Natural Language Processing, vol. 1, pp. 81–150. John Benjamins, Amsterdam (2000)

[Bunt, 1994] Bunt, H.C.: Context and Dialogue Control. Think 3, 19–31 (1994)

[Cassell et al., 1999] Cassell, J., Torres, O., Prevost, S.: Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. In: Wilks, Y. (ed.) Machine Conversations, pp. 143–154. Kluwer, The Hague (1999)

[Clark, 1996] Clark, H.H.: Using language. The Press Syndicate of the University of Cambridge (1996)

[Clark and Brennan, 1991] Clark, H.H., Brennan, S.E.: Grounding in Communication. In: Resnick, L.B., Levine, J., Teasley, S.D. (eds.) Perspectives on Socially Shared Cognition. American Psychological Association (1991)

[Fuhr et al., 1995] Fuhr, T., Socher, G., Scheering, C., Sagerer, G.: A Three-Dimensional Spatial Model for the Interpretation of Image Data. In: Proceedings of IJCAI 1995 Workshop on Representation and Processing of Spatial Expressions, Montreal, Canada (1995)

[Gapp, 1996] Gapp, K.-P.: Ein Objektlokalisationssystem zur sprachlichen Raumbeschreibung in dreidimensionalen Umgebungen: Formalisierung, Implementierung und empirische Validierung. PhD thesis, Department of Computer Science, Saarland University (1996)

[Gebhard, 2005] Gebhard, P.: ALMA—A Layered Model of Affect. In: Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems, Utrecht, The Netherlands, pp. 29–36 (2005)

[Göbel et al., 2006] Göbel, S., Schneider, O., Iurgel, I., Feix, A., Knöpfle, C., Rettig, A.: VirtualHuman: Storytelling and Computer Graphics for a Virtual Human Platform. In: Göbel, S., Spierling, U., Hoffmann, A., Iurgel, I., Schneider, O., Dechau, J., Feix, A. (eds.) TIDSE 2004. LNCS, vol. 3105, pp. 79–88. Springer, Darmstadt (2006)

[Heckmann, 2006] Heckmann, D.: Ubiquitous User Modeling. Akademische Verlagsgesellschaft. Berlin, Germany (2006)

[Kempe, 2005] Kempe, B.: Generation of Verbal and Nonverbal Utterances for Embodied Virtual Characters. Master's thesis, Department of Computer Science, Saarland University (2005)

[Kempe et al., 2005] Kempe, B., Pfleger, N., Löckelt, M.: Generating Verbal and Nonverbal Utterances for Virtual Characters. In: Proceedings of the International Conference on Virtual Storytelling 2005, Strasbourg, France, pp. 73–78 (2005)

[Knapp and Hall, 2002] Knapp, M.L., Hall, J.A.: Nonverbal Communication in Human Interaction. Wadsworth Publishing - ITP (2002)

[Levinson, 1983] Levinson, S.C.: Pragmatics. Press Syndicate of the University of Cambridge, Cambridge (1983)

[Levinson, 2003] Levinson, S.C.: Space in Language and Cognition. Press Syndicate of the University of Cambridge, Cambridge (2003)

[Löckelt, 2005] Löckelt, M.: Action Planning for Virtual Human Performances. In: Proceedings of the International Conference on Virtual Storytelling, Strasbourg, France (2005)

[Löckelt and Pfleger, 2006] Löckelt, M., Pfleger, N.: Augmenting Virtual Characters for more Natural Interaction. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 231–240. Springer, Heidelberg (2006)

[LuperFoy, 1991] LuperFoy, S.: Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions. PhD thesis, University of Texas at Austin (1991)

[Maass, 1996] Maass, W.: Von visuellen Daten zu inkrementellen Wegbeschreibungen in dreidimensionalen Umgebungen: Das Modell eines kognitiven Agenten. Dissertationen zur Künstlichen Intelligenz (DISKI). Akademische Verlagsgesellschaft, Berlin, Germany (1996)

[Pfleger, 2002] Pfleger, N.: Discourse Processing for Multimodal Dialogues and its Application in Smartkom. Diplomarbeit, Universität des Saarlandes (2002)

[Pfleger, 2004] Pfleger, N.: Context Based Multimodal Fusion. In: Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004), State College, PA, pp. 265–272 (2004)

[Pfleger and Alexandersson, 2006] Pfleger, N., Alexandersson, J.: Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems. In: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial), Postdam, Germany, pp. 2–9 (2006)

[Pfleger et al., 2003] Pfleger, N., Engel, R., Alexandersson, J.: Robust Multimodal Discourse. In: Proceedings of Diabruck: 7th Workshop on the Semantics and Pragmatics of Dialogue, Wallerfangen, Germany, pp. 107–114 (2003)

[Pfleger and Löckelt, 2006] Pfleger, N., Löckelt, M.: A Comprehensive Context Model for Multi-party Interactions with Virtual Characters. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 157–168. Springer, Heidelberg (2006)

[Pfleger and Schehl, 2006] Pfleger, N., Schehl, J.: Development of Advanced Dialog Systems with PATE. In: Processing of the International Conference on Spoken Language Processing (Interspeech/ICSLP), Pittsburgh, PA, pp. 1778–1781 (2006)

[Salmon-Alt, 2000] Salmon-Alt, S.: Interpreting Referring Expressions by Restructuring Context (Student Session). In: Proceedings of ESSLLI 2000, Birmingham, UK (2000)

[Wahlster and Kobsa, 1989] Wahlster, W., Kobsa, A.: User Models in Dialog Systems. In: Kobsa, A., Wahlster, W. (eds.) User Models in Dialog Systems, pp. 4–34. Springer, Berlin, Germany (1989)

[Yngve, 1970] Yngve, V.H.: On getting a word in edgewise. In: Papers from the Sixth Regional Meeting, pp. 567–577. Chicago Linguistics Society, Chicago (1970)

4
Social Intelligence in Virtual Groups

Rui Prada and Ana Paiva

IST-Technical University of Lisbon and INESC-ID
Avenida Prof. Cavaco Silva - Taguspark
2744-016 Porto Salvo, Portugal
{rui.prada,ana.paiva}@gaips.inesc.pt

Abstract. Autonomous virtual agents have the potential to promote social engagement of users in virtual environments, thus enhancing their interaction experience. This effect is supported by the interactions users and virtual agents perform together. These interactions are often in group scenarios, where both users and agents perform collaborative tasks.

However, in order to have successful group interactions, it is not enough to assure that the characters behave in a coherent manner from an individual perspective, it is also necessary that they exhibit behaviours that are coherent with the group's composition, context and structure.

Furthermore, nurturing social intelligence in virtual agents has been a major concern in the community of multi-agent systems, specially to endow them with skills to perform well in group tasks. However, when building agents to interact in group with users, to perform well in the task does not, necessarily, assure a good interaction experience. While, it is true that users expect the members of their group to perform well in the task, they also expect believable dynamics of the interpersonal relations.

In this chapter, we will present a model (SGD Model) to support the dynamics of group interactions that incorporates task related interactions as well as social emotional ones. The model defines the knowledge that each individual should build about the others and the group, and how this knowledge influences its action decision. We will also describe a study that was performed to assess the effect of the SGD Model in the interaction experience of users while playing a collaborative game with virtual counterparts.

4.1 Introduction

Autonomos synthetic characters are very useful to improve the interaction experience of users with virtual environments because they can foster the social context of such experience [7]. Therefore, they have been used in many different domains, for example, to improve educational aplications [24] [27] or to produce interesting entertainment experiences [18] [10]. They are particular important in computer games, since they constitute the main driving force to create sucessfull narrative experiences, which are important to improve gameplay [28] [17].

To create sucessful autonomous synthetic characters means to make them believable, thus, to give them the ability to create the "illusion of life" in the eyes of the viewers and lead them to the suspension of disbelief [8]. In other words,

N. Magnenat-Thalmann et al. (Eds.): New Adv. Virt. Hum.: Arti. Intel., SCI 140, pp. 113–132, 2008.
springerlink.com

autonomous synthetic characters must be coherent with the users' expectations. These expectations concern different issues. For example, users expect the characters to expresss a consistent personality, to reacting emotionally to important events and to behave according to the social context of the interaction.

The work here presented explore some of the issues of the influence of social context in the believability of autonomous synthetic characters. Namely, interactions in groups constituted simultaneously by users and autonomous members. The work is focused on groups with few members (small groups), that are committed to a collaborative task and that does not present a strong organizational structure. Thus, we are not concerned with groups as crowds or complex societies. In addition, the goal is to engage users, as well as the autonomous characters, as active members of the group.

The problem is that, usually, the autonomous characters lack the necessary social skills to sucessfully interact in group, for example, they are not able to exhibit behaviours that are coherent with the group's composition, context and structure. For this reason, their role in the group is, usually, very restricted and their autonomy is limited. For example, in Role Playing Games, where the user interacts with a group of several characters in order to solve the challenges of a fantasy world, the autonomous characters only take secondary roles, such as a salesperson, while the main characters are controlled by the user.

To tackle this problem we developed a model to support the behaviour of autonomous characters that interact in group (the SGD model) that allows each individual character to reason about the others and the group. This model was inspired by theories of group dynamics developed in human social psychological sciences and is driven by a characterization of the different types of interactions that may occur in the group, taking into account the socio-emotional interactions as well as the task-related ones.

To test the effects of the model in the interaction experience of users, it was implemented in the behaviour of autonomous synthetic characters that collaborate with a user in the resolution of tasks within a virtual environment in the context of a collaborative game called Perfect Circle. The game was used in a study that showed that the model had a positive effect on the users' trust and social identification with the group.

In this chapter we start by discussing some related work concerning the interaction of autonomous characters in group. Then, we present the SGD model and, afterwards, the Perfect Cicle game. Next, we describe the experiment that was conducted to assess the effects in the users' interaction experience and its results and finish with some conclusions.

4.2 Related Work

The problem of multiple autonomous synthetic characters that interact as a group in virtual environments has been previously addressed by several researchers. The first example of this can be found on Reynolds' Boids [25], which implements a flocking behaviour in a group of flying creatures (figure 4.1). In the

Fig. 4.1. Boids. A snapshot from Stanley and Stella in: Breaking the Ice.

same line of work we can additionally find research concerning the generation of crowds [22] that is often used in commercial systems for film creation. One well known example of this is “The Lord of the Rings” trilogy [23] that include numerous fighting scenes involving armies of thousands of warriors, the major part of these being played by synthetic actors.

The Boids’ flocking behaviour and crowd generation make use of an emergent group dynamics and result in a believable life-like group behaviour. However, characters in these examples do not have a deep social awareness and lack the ability to build social relations, which we believe to be essential for the interaction with a user.

Another example is the AlphaWolf system [31], which simulates the behaviour of a pack of six grey wolves (figure 4.2). In this system, the different synthetic characters are able to build domination-submission relationships. These relations are built in the form of emotional memories that drive the characters’ behaviour. In addition, three users can interact with the system and influence the behaviour of three of the wolves. According to the authors, AlphaWolf has successfully implemented a believable simulation of the group interactions in a pack of wolves, and has engaged the user in such interactions. However, users and the synthetic characters do not engage in the resolution of a collaborative task and do not have a strong notion of group.

Schmitt and Rist [29] developed a model of virtual group dynamics for small group negotiations (figure 4.3). In their system, users delegate the task of scheduling their appointment meetings to a virtual character. Characters will then meet in an arena and together negotiate the meetings’ times and dates. Each character has an individual personality and builds social attraction relations with the others. These relations and personality guide the characters’ interactions and support the generation of the negotiation dialogues. In the end, the dialogues are played for the users. The believability of the group dynamics is a key factor in this example as it supports the believability of the characters’

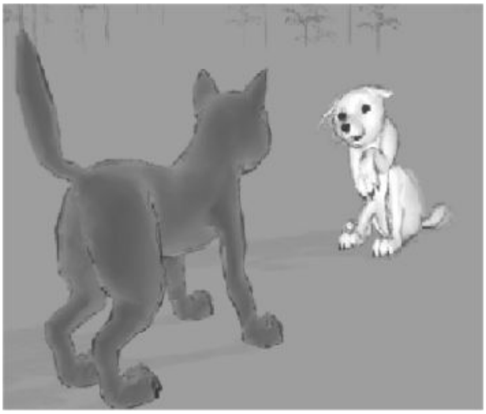


Fig. 4.2. AlfaWolf. The black pup is dominating the white pup.

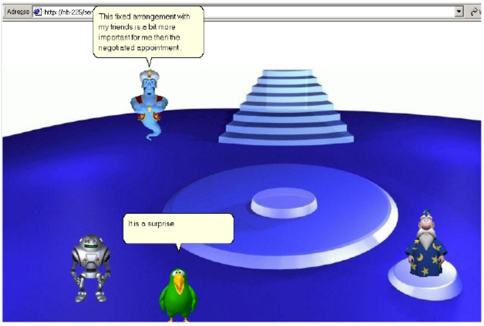

Adresse
http://nb-225/ser
This fixed arrangement with my friends is a bit more important for me than the negotiated appointment.
It is a surprise

Fig. 4.3. Snapshot of the Avatar Arena taken during the display of a generated negotiation dialogue

dialogues. But, users do not directly engage in the group interactions they are just played for them.

STEVE [27] is an example of a system where the users engage with a group of synthetic characters in a collaborative task (figure 4.4). It is used in a navy facility to train a team to handle possible malfunctions that may arise in a ship. The team can be composed of several human users and several virtual characters, which interact in a 3D virtual environment that simulates the ship and its equipment. In this scenario, the autonomous characters sucessfully engage in collaborative group activities with users, however, all interactions between the group members are related to the task and there is not the possibility for deeper social engagement.

In Demonstrate! [26] Rhalabi explore the use of emotions and personality to make the decions in a group, based on the Iterated Prisoner's Dilemma [5], in order to make the group more belieavable. The scenario explored represents a

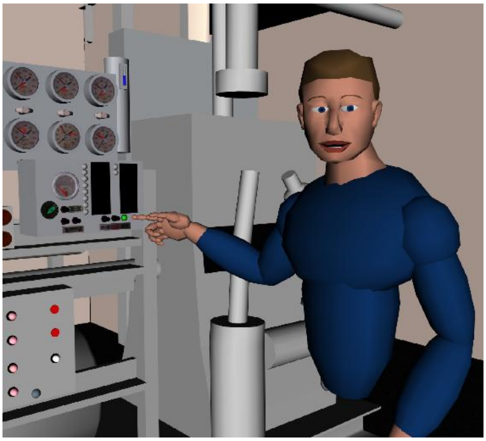


Fig. 4.4. STEVE pointing out a power light to a student

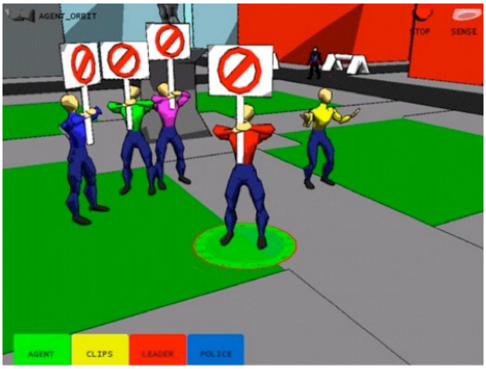


Fig. 4.5. A snapshot from Demonstrate!

gathering of people in the street to make a political demonstration (figure 4.5). The user controls the leader of this group and as s/he proposes an action, the autonomous charaters will choose to either cooperate or defect, based on their personality and emotions. This system, engages the user in a group of autonomous characters but has the limitation of not allowing the autonomous members to take the role of the leader.

Furthermore, Computer Role Playing Games (RPGs), such as "Star Wars: The Knights of the Old Republic"[9] (figure 4.6) or "Neverwinter Nights"[13], are systems that engage the users in a group with autonomous synthetic characters that perform a collaborative task. But, since the social interactions have an important role in this type of games, specially those that take place between the members of the group, and since the social skills of the autonomous characters are usually weak they only perform simple roles and are not deeply involved in

Fig. 4.6. A snapshot from “Star Wars: The Knights of the Old Republic”

the group task. Additionally, players frequently have some control over the autonomous characters, which reduces their autonomy. For example, in the ”Star Wars: Knights of the Old Republic” the player starts the adventure with one character, but, as the game evolves, other characters join the player’s quest and s/he will end up controlling simultaneously an entire party of several characters. This fact decreases the players’ perception of the synthetic members as individuals and increases the distance between the player and her/his character, which makes the players’ interaction experience in the group less interesting.

4.3 SGD Model: A Model for Group Believability

The model we propose to support the believability of autonomous characters’ behaviour in group, the SGD Model (Synthetic Group Dynamics Model), was build on the principle that each member of the group must be aware of the other members and the group itself. And, in addition, s/he should be able to build proper knowledge regarding the group’s social structure and to use this knowledge to drive her/his behaviour.

The group is modelled as a system composed of several autonomous agents that engage in interaction processes. These interactions create the dynamics of the system. They have effects on the group’s state and, at the same time, are influenced by that state. In other words, the preconditions for the occurrence of an interaction depends on the state of the group and the occurrence of an interaction will change, for example, the social structure of the group. For the definition of these interactions and their dynamics we have relied on theories of group dynamics developed in human social psychological sciences, in particular [11], [6] and [20].

In concrete terms, the the SGD Model was created as a module that influences the usual processes of a cognitive agent. Thus, the model influences the perception, knowledge building, behaviour and action processes of each agent (see figures 4.7).

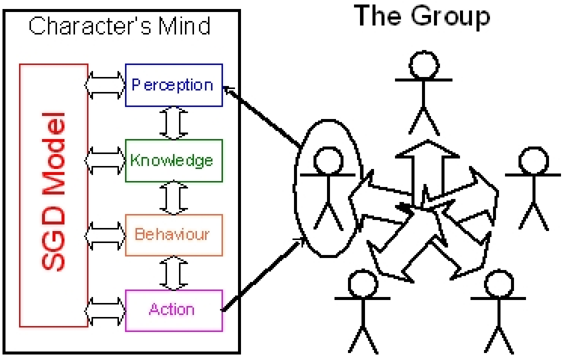


Fig. 4.7. The SGD Model in the mind of each member of the group

4.3.1 Target Groups

There are several definitions and types of groups. For that reason, we would like to clarify which kind of groups does the proposed model applies to, before entering in more details in its description.

As discussed before, our study is focused in groups that involve a human user with several synthetic characters. These groups perform in a virtual environment and their members are committed to solve collaborative tasks. Thus, the group interactions must evolve in such a way that make the resolution of those tasks possible.

In addition, the model applies to small groups, with only a few members, and without a strong organizational structure. We are not concerned with groups as crowds or complex organizations and societies of agents.

The members of the group are implemented as autonomous software agents that can engage into conversation and can manipulate objects in the virtual environment (e.g. get, give, use and drop items). The user is represented as an agent (avatar) in the system that is not autonomous but fully controlled by the user.

4.3.2 The SGD Model Components

The SGD Model is characterized in four different levels:

1. **the individual level** that defines the individual characteristics of each group member, such as their personality;
2. **the group level** that defines the group and its underlying structure;
3. **the interactions level** that defines the different classes of interactions and their dynamics;
4. **the context level** that defines the environment and the tasks that the agents can perform.

These four levels represent the knowledge that agents should build in order to implement the SGD Model in their behaviour. Furthermore, in addition to this knowledge, the agents' behaviour in the group relies on three processes:

1. **Classification of the Interactions:** the agent is aware of the actions in the group and classifies them into categories of interaction with specific semantics. For example, in this process the agent interprets if certain actions are helpful for the group of not. This process uses the information on the four levels of the agent's knowledge, specially on the interaction level, that defines the possible categories of interaction, and in the context level that defines how should the actions of the group be interpreted, for example, by means of social norms.
2. **Propagation of the Interaction Effects:** then, based on the identified category, the interaction produce some changes on the knowledge, in particular in the individual and group level. For example, the interaction may change the social relations established between the members that it engages.
3. **Influence of the Agent's Actions:** finally, the agent's perception of the group and its members influences the actions that if performs in the group. For example, if the agent is not motivated it will not try to solve the group's tasks.

4.3.3 The Individual Level

The individual level defines the knowledge that the agent build concerning the individual characteristics of each of the members of the group. This knowledge defines the members' abilities and their personality:

1. **The agent's abilities:** define the actions that each agent can perform in the environment associated with their levels of expertise (e.g. how good the agent is while performing each of these actions). The set of abilities is important to determine the agent level of expertise in the group, which is an important factor to define the agent's position in the group.
2. **The agent's personality:** we define the agent's personality using two of the dimensions proposed in the Five Factor Model [19]: *Extraversion* and *Agreeableness*. We only consider these two dimensions because they are associated with the ideas of dominant initiative and socio-emotional orientation proposed by Bales [1] while the other dimensions are more related to the task resolution which is not our main focus.
   a) *Extraversion:* is related to the dominant initiative of the agent. Thus, it will influence the agent's frequency of interaction.
   b) *Agreeableness:* is related to the socio-emotional orientation of the agent. It defines the type of socio-emotional interactions that the agent will favour. More agreeable agents will favour positive socio-emotional interactions, while less agreeable agents will favour negative socio-emotional interactions.

4.4 The Group Level

The group level defines the knowledge that the agents build concerning the group and its underlying structure, and additionally the agents' attitude towards the group.

First of all, the group is defined as a set of individuals that follows the definition presented in the previous section. But, more than just a set, the group is a unique and identifiable entity with an inherent structure. A group is defined by:

1. **The group identity:** identification is an important factor in the definition of a group. For that reason the group needs a unique name to allow it to be clearly distinct in the environment and enable the agents to recognize it and refer to it.
2. **The composition:** the composition is the set of individuals that are associated with the group. The composition may change over time as new members may be admitted or excluded.
3. **The structure:** the group structure is defined in different dimensions. According to Jesuino [16] the most common are the structure of communication, the structure of power and the structure of interpersonal attraction (sociometric structure [21]). As we are handling small groups the structure of communication should not be complex, since all characters may communicate directly with each other, thus, we decided not to include it in our model. The group structure is then defined in two dimensions: the *structure of power* that emerges from the members' social influence relations, and the *structure of interpersonal attraction* that emerges from the members' social attraction relations.

Furthermore, since the group's structure emerges from the social relations established between its members, the group characterization also depends on the definition of these social relations. Which, as said before, can be of two different types:

1. **Social attraction:** these relations define the interpersonal attraction of the members in terms of like (positive attraction) and dislike (negative attraction) attitudes. These relations are unidirectional and not necessarily reciprocal (e.g. if agent A has a positive attraction for agent B, this does not necessarily mean that agent B has a positive attraction for agent A).
2. **Social influence:** relations of influence define relations of power, they quantify the capacity of one agent to influence the behaviour of another. The influence is defined as the difference between the power that one individual can exert on another and the power that the other is able to mobilize to resist [14].

In addition, a member's social relations in conjunction with its level of expertise determine its position in the group. This position reflects the member's relative significance in the group which defines how important its contributions are and how well are they accepted by the group. For example, actions performed by members that have more social influence on the group members have stronger effects on the group process. Thus, the group position defines the agent's relative power in the group, which directly depends on (1) the overall social influence that the agent may exert on the others, (2) the attraction that the others feel for the agent and (3) the agent's relative expertise in the group.

4.4.1 The Interactions Level

The interaction level describes the knowledge that the agent builds concerning the group's interactions and their dynamics. This dynamics reflects, on (1) the changes that the group's interactions induce in the agent's perception of the group and, therefore, in the knowledge it builds about the group, and (2) in the rules that drive the behaviour of the agent in the group.

The central notion in the interactions level is the concept of *interaction*. An *interaction* occurs when agents execute actions that can be perceived and evaluated by others. In fact, it may consist on several actions that are performed in a certain pattern. These actions can be performed simultaneously, which means that more than one agent may be involved in the same *interaction*. In addition, other agents may support the *interaction* but not be directly involved in its execution. For example, agents may agree with a certain *interaction* and explicitly show their support for its execution without performing a single action concerning the *interaction* other than the declaration of support.

Moreover, each *interaction* has a certain strength in the group that defines its relative importance in the group process. Additionally, each *interaction* may affect only certain members of the group. For example, when a member of the group encourages another to perform a task, the effects of the encouragement will only be reflected on the agent that was encouraged. The strength of an *interaction* in the group is directly related to the position in the group of the *interaction*'s performers and supporters.

The Classification of the Interactions. In order to model the dynamics of the group process we have divided the several possible group interactions into different categories. This categorization is then embedded in the knowledge that the agent has a priori. It will support the agent's process of perception and identification of the interactions when it asserts new interaction facts in its knowledge base.

Furthermore, although the interaction is closely related to the actions that the agents perform, its classification is more than just the classification of the actions themselves. It depends on the actions' results, on the context of the execution, and also on the agents' perception of the group. Thus, for example, the same action can be perceived as a positive interaction to the group by an agent but as a negative by another.

The classification that the SGD Model presents was based on the categories that Bales proposed on his IPA system [6]. Thus, it follows the same main distinction between socio-emotional and instrumental interactions, and divides the interactions into positive and negative.

On the socio-emotional level we use six categories similar to those presented by Bales. We consider three positive socio-emotional interactions (*agree, encourage and encourage group*) and three negative social emotional interactions that are opposed by symmetry (*disagree, discourage and discourage group*).

- Positive socio-emotional interactions
    1. **Agree:** This class of interactions show the support and agreement of an agent towards one of the interactions of another agent consequently raising the importance of that interaction in the group.
    2. **Encourage:** These interactions represent an agent efforts to encourage another agent, consequently facilitating its social condition (e.g. increasing its motivation).
    3. **Encourage Group:** This class of interactions are similar to those in the *Encourage* category but apply to the group itself. These interactions encourage the group and facilitate the group's social structure (e.g. its cohesion).
- Negative socio-emotional interactions
    1. **Disagree:** This class of interactions show disagreement of an agent towards one of the interactions of another agent, consequently decreasing the importance of that interaction in the group.
    2. **Discourage:** These interactions represent an agent's hostility towards another agent and its efforts to discourage it.
    3. **Discourage Group:** This class of interactions are similar to those in the *Discourage* category but apply to the group itself. These interactions discourage the group and raise the entropy of its social structure.

The categories proposed by Bales at the instrumental level focus mainly on speech acts. And, in addition, there is not a clear connection between the instrumental interactions an the task itself. However, in the context of virtual environments, the interactions that are not based on speech acts are very important because the agents may manipulate the objects defined in the environment. Also, the design of the interactions' influence on a problem solving group and its members is easier if the interactions' definition is based on the concept of "problem". Therefore, following these two principles we defined four instrumental interactions: two positive (*facilitate problem, gain competence*) and two negative (*obstruct problem, loose competence*), that do not have a direct correspondence in the IPA instrumental categories.

- Positive instrumental interactions
    1. **Facilitate Problem:** This class of interactions represents the interactions of an agent that solves one of the group's problems or facilitates its resolution.
    2. **Gain Competence:** These interactions make an agent more capable of solving a problem. This includes, for example, the learning of new capabilities or the acquisition of information and resources.

- Negative instrumental interactions
  1. **Obstruct Problem:** This class of interactions represents the interactions of an agent that complicates one of the group's problems or makes its resolution impossible.
  2. **Loose Competence:** These interactions make an agent less capable of solving one problem. For example, by forgetting information or loosing the control of resources.

The Interactions' Dynamics. As stated before, the interactions create the dynamics in the group. Such dynamics are supported by the classification presented in the previous section 4.4.1 and are modelled by a set of rules that follow the ideas found in the social psychological theories of group dynamics, like for example, the theory of social power by French and Raven [14] and Heider's balance theory [15]. These rules define, on one hand, how the agent's and the group's state influence the occurrence of each kind of interaction and, on the other hand, how the occurrence of each type of interaction influences the agent's and group's state.

During the group process, each member observes the actions that are being executed by the others and tries to identify patterns that match each of the proposed categories. This classification is done according to the current context and depends on the individual view of each member. Thus, for example, if two members have different views concerning the group's tasks, some actions may be perceived by one member as helpful to the resolution of these tasks and, therefore, classified as *FacilitateProblem* but can be perceived by the other as disadvantageous and, therefore, classified as *ObstructProblem*.

Furthermore, when members identify the occurrence of one interaction, they react to it according to the classification that they internally gave to the interaction. These reactions are translated into changes on the perceived knowledge of the group, specially in its structure. *Instrumental* interactions are related to changes in the relations of *social influence*, thus, each member that is responsible for positive *instrumental* interactions will raise her/his *influence* over the others and will decrease it in the case of a negative *instrumental* interactions. In turn, *socio-emotional* interactions are related to changes in the relations of *social attraction*, thus, each member that is target of a positive *social-emotional* interaction will raise her/his *attraction* for the performers and will decrease it in case of a negative *social-emotional* interaction. The *motivation* of the members involved in the interaction may also improve in the case of positive interactions and decrease otherwise. These rules are resumed in table 4.1.

Moreover, in order to keep the social relations balanced [15], the *social-emotional* interactions may have effects on a member of the group this is not directly involved in the interaction. For example, imagine that John is encouraging Frank because he failed to perform a certain task and Mary observed this event. Mary knows that Frank will increase his social attraction for John and this will lead to changes in her own relation with the two. For instance, if Mary has a positive relation with Frank then her relation with John may improve. But,

Table 4.1. The effects of the interactions on motivation (Mot), social influence (SI) and social attraction (SA). P denotes the member that performs the interaction and T the target of the interaction. The symbols in the table define if the value increases of decreases.

Interaction	**Mot(P)**	**Mot(T)**	**SI(P,T)**	**SA(T,P)**
Pos-Instr(P,T)	+		+	
Neg-Instr(P,T)	-		-	
Pos-SocEmot(P,T)		+		+
Neg-SocEmot(P,T)		-		-

Table 4.2. The effects of the interactions on the social attraction of an observer. The values on the table reflect the changes on SA(O,P). The symbols in the table define if the value increases of decreases.

Interaction	**SA(O,T) $>$ 0**	**SA(O,T) $<$ 0**
Pos-SocEmot(P,T)	+	-
Neg-SocEmot(P,T)	-	+

Table 4.3. The influence on the four major categores of interactions: Socio-emotional positive (SE-Pos), Socio-emotional negative (SE-Neg), Instrumental positive (I-Pos) and Instrumental negative (I-Neg). The symbols in the table define if the value increases of decreases.

Variable	**SE-Pos**	**SE-Neg**	**I-Pos**	**I-Neg**
Motivation(P)	+	+	+	+
Extraversion(P)	+	+	+	+
GroupPosition(P)	+	+	+	+
Agreeableness(P)	+	-		
GroupPosition(T)	+	-		
Influence(P, T)	-	+		
Influence(T, P)	+	-		
Attraction(P, T)	+	-		
Skills(P)			+	-

if, on the other hand, she has a negative relation with Frank then her relation with John may become worse. Table 4.2 resumes these rules.

The intensity of the interactions' effects depends on the *position* in the group of the members that perform them. Thus, for example, encouragements performed by members with a better *position* will increment more the target's *motivation*.

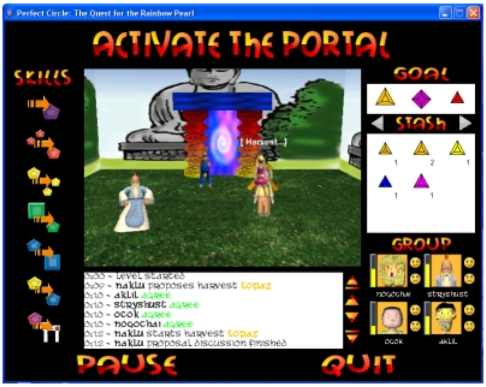
The knowledge built regarding the group, in its three different levels, regulate the behaviour of the member in the group. This is reflected in a set of conditions that determine the frequency of occurrence of each type of interaction. These conditions depend on individual characteristics, such as *motivation* and *personality*, and on the social structure of the group [30] [20] [1].

Table 4.3 resumes the influence of each of these variables regarding the four main categories of interaction. For example, the first three lines express the general rules for the frequency of all types of interaction, which state that: *"highly motivated agents engage in more interactions, as well as agents with a good position in the group or high extraversion"*. Another example, concerning the social relations, is expressed in line 7, which states that: *"a character will engage in more positive socio-emotional interactions towards members that have influence over him"*. Note that decisions are probablistic. The abovementioned rules only suggest the frequency of interaction. Thus, for example, a less motivated agent can perform tasks, but not very often.

4.5 The Perfect Circle Game

To test the effects of the SGD Model we have developed a collaborative game called "Perfect Circle: the Quest for the Rainbow Pearl"[1] and implemented the SGD Model in the mind of the autonomous characters of that game.

The game takes the user into a fantasy world where he joins a group of four other characters to search the world for a magic item. To achieve this, the group must travel around the world through magic portals that are activated by the powers of some gemstones. Their task is to gather and manipulate the gemstones in order to get the required ones that will open the portal (see figure 4.8). To achieve this, characters need to apply their individual abilities in order to change the gems' form, size and colour. For example, if the group has two small rubies but it needs one medium sized ruby, one character can use its ability to merge the small stones into a bigger one. In addition, two or more characters can combine their efforts if they all have the same ability. As a result, the probability of success of the action becomes higher.



```
0:00 - level started
0:09 - naklu proposes harvest topaz
0:10 - aklil agree
0:10 - stryshust agree
0:10 - ocok agree
0:10 - nogochai agree
0:12 - naklu starts harvest topaz
0:12 - naklu proposal discussion finished
```

Fig. 4.8. The group is trying to activate one of the portals in order to move further

Furthermore, every character in the group is engaged in the same goal, thus trying to solve the same task. However, there are many ways to reach a solution, and if each of the characters follows its own, the group may never solve the task. Thus, characters have to coordinate their actions in order to follow a similar strategy in the search for the correct gems to activate the portal.

[1] This game can be downloaded from *http://web.tagus.ist.utl.pt/˜ rui.prada/perfect-circle/*

For this reason, every action that is performed in the group, concerning the resolution of the task, is discussed by the group beforehand. The discussion protocol has three different steps:

1. First, one character declares that s/he wants to take a certain action (e.g. *"I think that it will be best if I merge these two sapphires"*).
2. The other characters can respond to the proposal with one of the following: (1) *Agree* with the course of action; (2) *Join* the action and help in the execution; (3) *Disagree* with the course of action.
3. Then, based on the opinions expressed by the group, the character decides to proceed with the execution of the action or to withdraw the proposal. If s/he decides to proceed with the action then s/he starts its execution. All other characters that have decided to join the action start their contributions to the joint execution.

Furthermore, the group interactions are not restricted to the execution of the task. Each member can, at any time, engage in social-emotional interactions by encouraging or discouraging the other members of the group. For example, if one character just messed up the task the others can discourage her/him and tell her/him to stop trying or, on the other hand, may encourage her/him tell her/him that if was just bad luck.

4.6 Study

To evaluate the effects of the SGD Model in the interaction experience of users we have conducted an experiment was at our university with 24 students of computer engineering, being 20 of them male and 4 female. The subjects' age ranged between 19 and 31 years old.

4.6.1 Independent Variables

The experiment was conducted with two main independent variables: the use of the model SGD Model to convey the believable group dynamics and the initial structure, and consequent cohesion level, of the group.

1. **The Use of the model:** SGD Model two different versions of the game were built: one where the characters followed the model SDG Model and other where they did not. When the characters did not use the model they were not able to engage in socio-emotional interactions, except *Agree* and *Disagree* (without any socio-emotional connotation). In addition, their frequency of interaction was always constant and the decision to proceed with a proposed action was not weighted by the members' position in the group, it was a simply majority rule.
2. **The Group's Initial Structure:** subjects can start the game in a group with non neutral initial social relations, which means that the initial group can have levels of cohesion that may be either very high or very low. Two

different scenarios were considered: one where the group had neutral social relations and a second one where the members of the group disliked each other, which, took the group cohesion to very low levels. Note that this condition could only be applied when the game was run with the believable group dynamics model.

4.6.2 Dependent Variables

To assess the quality of the subjects' interaction experience while playing the game we have measured their satisfaction with the game as well as their trust and social identification with the group, since, according to Allen et al. [2] these two variables are related to the satisfaction of people when interacting in group. Thus, the three dependent variables are:

1. **Group Trust:** people's trust on a group has a positive effect on their perceptions about their experience in the group [12], which consequently leads to a more satisfactory interaction [3].
2. **Group Identification:** according to Ashforth and Mael [4] social identification is, in addition to social trust, one of the factors that foster the members of a group to be more engaged and more satisfied with the group.
3. **Satisfaction with the Game:** computer games are supposed to be fun, thus, the user should enjoy every moment that s/he spends with the game. Hence, to improve the interaction experience, as stated in the initial hypothesis, would imply also to increase the user's fun.

4.6.3 Procedure

The experiment was divided into four sessions of two hours each. In each session we had six different participants each on a different computer with the *Perfect Circle* game installed. The game was installed according to three different conditions (two computers for each condition):

(C1) In the first condition the game was installed without our model for believable group dynamics.
(C2) In the second condition the game was installed with the model and the group had neutral social relation in the beginning of the game.
(C3) In the third condition the game was installed with the model but the members of the group started with negative social attraction relations, thus, the level of cohesion of the group was very low.

Furthermore, apart from the differences mentioned, all the other details were similar for the three conditions. The four autonomous characters had the same name, the same appearance, the same personality and the same skills. In addition, the sequence of the game puzzles was predefined and the same for all the subjects. This sequence was randomly generated beforehand. The subjects were selected on the fly in the beginning of each session and they chose freely which computer to use. Subjects played the game for one and a half hour and at the end were requested to fill a questionnaire.

This process was repeated in the four sessions, which in the end gave a sample of eight subjects for each of the conditions.

4.6.4 Results

Concerning the subjects' trust in the group we have reached some significant results. As shown in figure 4.9, the subjects who played the game with the SGD Model had more trust in the group than those who played without the model. Furthermore, the fact that the group was initially cohesive, or not, did not influence the final levels of trust.

Cond.	N	Mean Rank
1	8	7,31
2	8	15,19
3	8	15,00

Chi-square: 6,492; *Asymp. Sig:* ,039

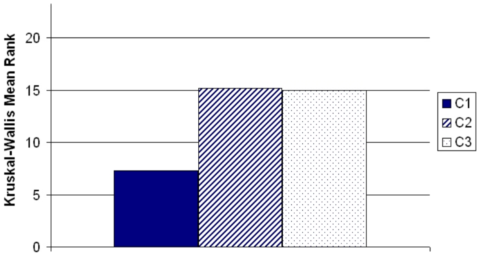


Fig. 4.9. The subjects' trust in the group across the three different conditions of the experiment

The results regarding the subjects' identification with the group are similar to those verified concerning the group trust. Which means that the SGD Model also had effect on the subjects' social identification with the group. Figure 4.10 shows these results and, as we can see, the identification with the group is higher in the two conditions where the synthetic characters used the SGD Model to drive their behaviours.

In addition, there are differences in the social identification in relation to the initial group cohesion. It seems that the most cohesive group induced lower levels of identification in the subjects. We believe that this effect may be related to the fact that the socio-emotional interactions in the highest cohesion group are essentially positive, which is probably less believable than a scenario where both, positive and negative, socio-emotional interactions occur, as in the case of the third condition.

Concerning the subjects' satisfaction with the game, we reached some interesting results. As figure 4.11 shows, the general satisfaction was the highest in

Cond.	N	Mean Rank
1	8	8,00
2	8	12,94
3	8	16,56

Chi-square: 5,960; *Asymp. Sig:* ,051

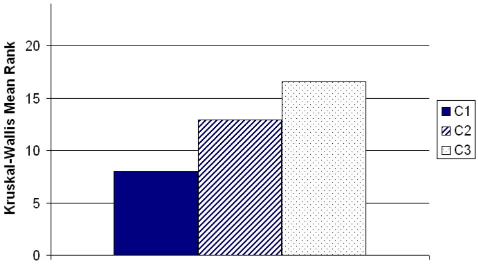


Fig. 4.10. Subjects' identification with the group across the three different conditions of the experiment

Cond.	N	Mean Rank
1	8	12,94
2	8	8,56
3	8	16,00

Chi-square: 4,503; *Asymp. Sig:* ,105

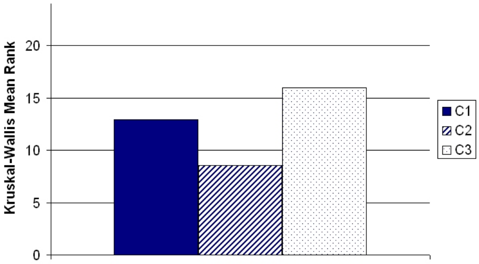


Fig. 4.11. Subjects' general satisfaction with the game across the three different conditions of the experiment

the case of the third condition but it was the lowest in the case of the second condition.

This effect is surprising as it contradicts, in a certain way, the other results, since the effects of the SGD Model were always positive in the case of the other two variables. Nevertheless, in the third condition, where the group was initially non cohesive, the positive effect still applies to the satisfaction with the game. Thus, there was a particular element of the second condition that did not please the players. Our hypothesis is that, since the socio-emotional interactions in a cohesive group are more likely to be positive, the subjects did not find the group itself to be challenging, and therefore were more bored with the group interactions. However, with this study we can not confirm this hypothesis with confidence.

4.7 Conclusions and Future Work

In this chapter we argued that group believability of synthetic characters is important, when among the group, we have characters and users interacting with each other, which is often the case of virtual environments. Then, to achieve such group believability, we have proposed a model inspired by theories of group dynamics developed in human social psychological sciences. The dynamics is driven by a characterization of the different types of interactions that may occur in the group. This characterization addresses socio-emotional interactions as well as task related interactions.

This model was successfully used in the context of a collaborative game, and the experiment conducted in that scenario demonstrated the positive effect that the model can have on the users' interaction experience, specially on the users' trust and identification with the synthetic group.

Furthermore, we found some evidence that if the group starts with low levels of the cohesion players have more fun playing the game. This gives us some evidence that players might prefer playing in groups that have higher level of conflict.

References

1. Acton, S.: Great ideas in personality - theory and research (last access on January 2005), `http://www.personalityresearch.org/bigfive.html`
2. Allen, K., Bergin, R., Pickar, K.: Exploring trust, group satisfaction, and performance in geographically dispersed and co-located university technology commercialization teams. In: Proceedings of the NCIIA 8th Annual Meeting: Education that Works, March 18–20 (2004)
3. Ang, J., Soh, P.H.: User information satisfaction, job satisfaction and computer background: an exploratory study. Information and Management 32, 255–266 (1997)
4. Ashforth, B.E., Mael, F.A.: Social identity and the organization. Academy of Management Review 14, 2–39 (1989)
5. Axelrod, R.: The Evolution of Cooperation. Basic Books, Inc., New York (1984)
6. Bales, R.F.: Interaction Process Analysis. The University of Chicago Press, Chicago (1950)
7. Bates, J.: Virtual reality, art and entertainment. Presence: teleoperators and Virtual Environments (1992)
8. Bates, J.: The role of emotions in believable characters. Communications of the ACM 37(7), 122–125 (1994)
9. Bioware. Star wars: Knights of the old republic (2003), `http://www.bioware.com/games/knight_sold_republic/`
10. Blumberg, B. (void*): A cast of characters. In: Visual proceedings of the conference on SIGGRAPH 1999, p. 169. ACM Press, New York (1999)

11. Cartwright, D., Zander, A.: Group Dynamics: research and Theory. Harper and Row, New York (1968)
12. Driscoll, J.W.: Trust and participation in organizational decision making as predictors of satisfaction. Academy of Management Journal 21, 44–56 (1978)
13. O. Entertainment. Neverwinter nights 2 (2006), `http://www.atari.com/nwn2`
14. French, J.R.P., Raven, B.H.: Bases of Social Power. Group Dynamics: Research and Theory. Harper and Row, New York (1968)
15. Heider, F.: The Psychology of Interpersonal Relations. Wiley, New York (1958)
16. Jesuno, J.C.: Estrutura e processos de grupo: Interaces e factores de eficcia. In: Psicologia Social. Fundao Calouste Gulbenkian (2000)
17. Lindley, C.A.: The gameplay gestalt, narrative, and interactive storytelling. In: Proceedings of Computer Games and Digital Cultures Conference, pp. 203–215 (2002)
18. Martinho, C., Paiva, A.: Pathematic agents. In: Proceedings of the 3rd International Conference on Autonomous Agents, Seattle, USA. ACM Press, New York (1999)
19. McCrae, R., Costa, P.: Toward a new generation of personality theories: theoretical contexts for the five factor model. In: The five factor model of personality: Theoretical perspectives, pp. 51–87. Guilford, New York (1996)
20. McGrath, J.E.: Groups: Interaction and Performance. Prentice-Hall, Englewood Cliffs (1984)
21. Moreno, J.L.: Who Shall Survive? Nervous and Mental Disease. Publishing Co., Washington D.C (1934)
22. Musse, S.R., Thalmann, D.: Hierarchical model for real time simulation of virtual human crowds. IEEE Transactions on Visualization and Computer Graphics 7(2), 152–164 (2001)
23. New-Line-Productions. The lord of the rings movies official site (2001), `http://www.lordoftherings.net`
24. Prada, R., Machado, I., Paiva, A.: Teatrix: virtual environment for story creation. In: Proceedings of the Fifth International Conference on Intelligent Tutoring Systems, Montreal, Canada. Springer, Heidelberg (2000)
25. Reynolds, C.W.: Flocks, herds, and schools: a distributed behavioural model. In: Computer Graphics (SIGGRAPH 1987 Conference Proceedings), pp. 25–34 (1987)
26. Rhalibi, A.E., Baker, N., Merabti, M.: Emotional agent model and architecture for npcs group control and interaction to facilitate leadership roles in computer entertainment. In: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology, pp. 156–163. ACM Press, New York (2005)
27. Rickel, J., Johnson, W.L.: Virtual humans for team training in virtual reality. In: Proceedings of the Ninth International Conference on AI in Education, pp. 578–585 (1999)
28. Schell, J.: Understanding entertainment: story and gameplay are one. Computers in Entertainment 3(1) (2005)
29. Schmitt, M., Rist, T.: Avatar arena: virtual group-dynamics in multi-character negotiation scenarios. In: 4th International Workshop on Intelligent Virtual Agents, p. 358 (2003)
30. Shaw, M.E.: Group Dynamics: the Psychology of Small Group Behaviour. McGraw-Hill, New York (1981)
31. Tomlinson, B., Blumberg, B.: Social synthetic characters. Computer Graphics 26(2) (May 2002)

5 Modeling Gesticulation Expression in Virtual Humans

Celso M. de Melo[1] and Ana Paiva[2]

[1] USC, University of Southern California
demelo@usc.edu
[2] IST – Technical University of Lisbon and INESC-ID,
Avenida Prof. Cavaco Silva – Taguspark,
2780-990 Porto Salvo, Portugal
ana.paiva@inesc-id.pt

Abstract. Gesticulation is the kind of unconscious, idiosyncratic and unconventional gestures humans do in conversation or narration. This chapter reviews efforts made to harness the expressiveness of gesticulation in virtual humans and proposes one such model. First, psycholinguistics research is overviewed so as to understand how gesticulation occurs in humans. Then, relevant computer graphics and computational psycholinguistics systems are reviewed. Finally, a model for virtual human gesticulation expression is presented which supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientation palm axis, orientation angle and handedness) and dynamic features; (b) synchronization between gesticulation and synthesized speech; (c) automatic reproduction of annotations in GestuRA, a gesticulation transcription algorithm; (d) expression control through an abstract integrated synchronized language – Expression Markup Language (EML). Two studies, which were conducted to evaluate the model in a storytelling context, are also described.

5.1 Introduction

Humans express thought through gesticulation. Gesticulation is the kind of unconscious, idiosyncratic and unconventional gestures humans do in conversation or narration. They tend to focus on the arms and hands, though other body parts may be involved. Furthermore, gesticulation and speech, which are believed to be different sides of the same mental process, co-express the same underlying idea unit and synchronize at various levels. [1,2]

The problem of modeling gesticulation can be divided into the sub-problems of generation and execution. *Gesticulation generation* concerns with the simulation of the speech and gesture production process, i.e., the distribution of communicative intent across modalities and selection of proper surface realizations which, in the case of gestures, correspond to constraints on static and dynamic features of the arms and hands. *Gesticulation execution* is more akin to the body and concerns with the actual animation, in a synchronized fashion, of the static and dynamic constraints which define the gesture. This chapter will focus on the latter but, overview the former.

To clarify the challenges involved in this endeavor, a virtual human model for gesticulation expression is described. The model supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language

N. Magnenat-Thalmann et al. (Eds.): New Adv. Virt. Hum.: Arti. Intel., SCI 140, pp. 133–151, 2008.
springerlink.com

hand shapes, orientations and positions) and dynamic features; (b) multimodal synchronization, including between gesticulation and speech; (c) automatic reproduction of annotated gesticulation according to *GestuRA*, a gesture transcription algorithm; (d) expression control through a markup integrated synchronized language.

The rest of the text is organized as follows: section 5.2 overviews gesticulation research in psycholinguistics; section 5.3 describes relevant computational models; section 5.4 presents a gesticulation expression model for virtual humans as well as an evaluation study; finally, section 5.5 draws conclusions and discusses future work.

5.2 Gesticulation and Psycholinguistics

Human gestures can be categorized into three subclasses [3]: gesticulation; emblems; and signs. *Emblems* are culturally dependent gestures which have conventionalized meaning. An example is the American V (of victory) gesture, executed with the palm facing the listener. *Sign languages* consist of communication languages expressed through visible hand gestures. Examples are languages used by the deaf, such as the Portuguese Sign Language [4]. Finally, *gesticulation*, which is the focus of this chapter, is the kind of idiosyncratic, unconventional and unconscious gestures humans do in narrations or conversations [1,2]. Gesticulation is tightly synchronized with speech, is structured in phases and can be interpreted according to several dimensions.

5.2.1 Gesticulation and Speech

Gestures which occur when a person is speaking manifest verbal thought. Verbal thought, which does not include all forms of thought, nor all forms of speech, is the kind of thought which resides in the intersection between thought and speech. It is believed that speech and gesticulation are manifestations of the same underlying process [1,2]. Thus, gesticulation and speech co-express the same underlying idea unit possibly in non-redundant ways, as they synchronize at the semantic and pragmatic levels, develop together in childhood and deteriorate together in aphasia. Through gesticulation, however, information is conveyed in a fundamentally different way than through speech: (a) gesticulation is not combinatoric – two gestures produced together do not combine to form a larger one with a complex meaning; (b) there is no hierarchical structure in gesticulation as in language; (c) gesticulation does not share the linguistic properties found on verbal communication.

5.2.2 Gesticulation Structure

According to how it unfolds in time, gesticulation can be structured hierarchically into units, phrases and phases [5,6]. A *unit*, which is the highest level in the hierarchy, is the time interval between successive rests of the limbs. A unit may contain various phrases. A *phrase* is what is intuitively called ‘gesture’ [2]. A phrase consists of various *phases*: (a) preparation, where the limbs position themselves to initiate the gesture; (b) pre-stroke hold, where a hold occurs just before the stroke; (c) stroke, which is the only obligatory phase, where actual meaning is conferred. The stroke is synchronous with its co-expressive speech 90% of the time [7] and, when asynchronous, precede the semantically related speech; (d) post-stroke hold, where a hold occurs after the stroke, before initiating retraction; (e) retraction, where the limbs return to the resting position. Preparation, stroke and retraction were introduced by Kendon [8] and the holds by Kita [9].

5.2.3 Gesticulation Dimensions

McNeill and colleagues characterize gesticulation according to four dimensions [1,2]: (1) *iconicity*, which refers to gesticulation features which demonstrate through its shape some characteristic of the action or event being described; (2) *metaphoricity*, which is similar to iconics however, referring to abstract concepts; (3) *deixis*, which refers to features which situate in the physical space, surrounding the speaker, concrete and abstract concepts in speech; (4) *beats*, which refer to small baton like movements that do not change in form with the accompanying speech. They serve a pragmatic function occurring, for instance, with comments on one's own linguistic contribution, speech repairs, and reported speech. According to McNeill ([2], p.42), "multiplicity of semiotic dimensions is an almost universal occurrence in gesture". Thus, it makes more sense to speak of dimensions and saliency rather than exclusive categories and hierarchy.

5.2.4 Gesticulation Models

Several gesticulation production models have been proposed. McNeill's *growth point model* [1,2] explains verbal thinking through growth points, which represent idea units. In a growth point two unlike modes of thinking – linguistic and imagistic – are active and this instability resolves by accessing stable language forms and materializing into gesture. This materialization increases with the unpredictability of the idea unit, i.e., with its opposition to current context. In contrast, extending Levelt's *speaking* model [10], various modular information processing models have been proposed including by de Ruiter, Krauss and Kita & Özyürek. In *de Ruiter's sketch model* [11] the conceptualizer – which transforms communicative intent into a propositional form called the preverbal message – receives as input communicative intent and outputs a *sketch* holding gesture form specifications. These specifications rely on a *gestuary* which stores predefined gesture templates which impose constraints on features. Synchronization is achieved through signal passing between modules. Krauss's model [12], contrasting to de Ruiter's and McNeill's assumption of imagistic knowledge, is a *featural model*, i.e., concepts are represented as propositional and non-propositional (visuospatial) features. During gesture production, a subset of the non-propositional features is selected to pass down to a motor planner which generates form. The model also describes gesture effects on lexical retrieval. As in de Ruiter's model, synchronization is achieved through signal passing. Kita and Özyürek [13] propose a model which says, contrasting to Krauss' and de Ruiter's models, that gestures are influenced by the speaker's language.

5.2.5 Implications for Computer Science

The psycholinguistics research presented in this section leads to several requisites for a computational model of gesticulation:

- *Gesticulation should, at least, span arms and hands*, as it tends to focus in these body parts;
- *Gesticulation and speech should be able to synchronize at the sub-second time granularity*, as they are believed to be different sides of the same underlying mental process and synchronize at the semantic and pragmatic levels;
- *It should be possible to describe gesticulation at the phase level*, as they distinguish parts which are motivated by physical, synchronization or meaning constraints. Phases are also crucial for gesture fusion in co-articulation effects;
- *Gesticulation can be described through constraints on its features*, in concrete, as sequences of static (hand shape, orientation and position) and dynamic constraints (motion profiles). The feature-based approach is justified for several reasons. First, describing gesticulation according to dimensions and saliency suggests that meaning distributes across the affordances of the upper limbs and hands and thus, rather than overall form a more granular (or feature-based) description is possible. Second, a feature-based approach is compatible with most speech and gesture production models: the imagistic component in McNeill's growth points ultimately materializes into gesture features; de Ruiter's sketch model revolves around the concept of gesture templates (in a gestuary) which correspond to constraints on features; Krauss actually considers knowledge representation as feature-based; finally, Kita & Özyürek even though not detailing gesture morphology, motivate their model with motion gestures described according to features.

5.3 Gesticulation and Computer Science

Building a gesticulation expression computational model comes with many challenges, Fig. 5.1. First, it is necessary to build a *virtual human* which has a body which can be animated to gesticulate. This challenge is in the domain of computer graphics. Second, it is necessary to solve the *gesticulation execution* problem, which concerns with animating a gesticulation plan. Third, it is necessary to solve the *gesticulation production* problem, which isn't independent of speech production and concerns with converting communicative intent into synchronized verbal and gesticulation plans. Building on virtual human models, computational psycholinguistics systems address these last two issues. Finally, *interfaces* should be built between these layers to promote modularity. In

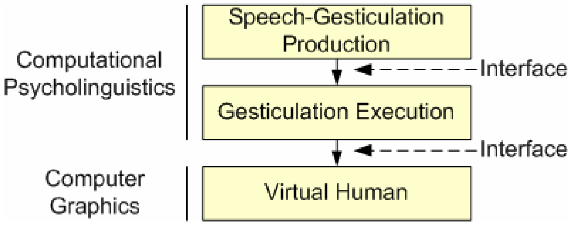


Fig. 5.1. A framework for gesticulation expression

this regard, several markup languages have been proposed. Section 5.4 describes one approach which focuses on the gesticulation execution problem.

5.3.1 Gesticulation and Computer Graphics

In its simplest form, building a virtual human consists of defining a hierarchical skeleton and a mesh for the skin. Animating the skeleton leads to skin mesh deformation using the vertex blending technique [14]. Several animation mechanisms have been explored [2]: (a) motion capture, where animation is driven by a human actor; (b) keyframe animation, where a human artist defines keyframes and in-between frames are automatically generated; (c) inverse kinematics, where animation of the body's extremities automatically animate the rest of the chain; (d) dynamics-based animation, which generates physically realistic animation.

Particularly relevant to gesticulation are specialized models of the hands. Thompson [16] proposes an anatomically accurate hand model based on tomographic scans. Wagner [17] argues for individual models of the hand by comparing the anthropometry of pianists with regular people. Moccozet [18] proposes a three-layer hand model – skeleton, muscle and skin – where Dirichlet free-form deformations are used to simulate muscle and realistic skin deformation. Sibille [19] proposes a real-time generic anatomical hand model. Hand motion is based on dynamics, mass-spring meshes are used to calculate soft tissue deformations and, finally, the system handles collision detection. Finally, Albrecht [20] also proposes a real-time anatomical human hand model. Motion relies on a realistic muscle model based on anatomical data, mechanical laws and a mass-spring system. Even though these models have great potential to generate realistic gesticulation, thus far, most computational psycholinguistics systems have used far simpler hand models.

5.3.2 Gesticulation and Computational Psycholinguistics

Building on the aforementioned graphic models, several computational psycholinguistics systems have been proposed to address the gesticulation production and execution problems. Animated Conversation [21], developed by Cassell and colleagues, is a rule-based system capable of synchronizing gestures of the right type with co-occurring speech. *Real Estate Agent (Rea)* [22,23] presents an embodied conversational agent capable of proper distribution and realization of communicative intent across speech and gesture. Cassell et al. [24] also propose the *Behavior Expression Animation Toolkit (BEAT)* which receives as input text and, based on rules, automatically generates appropriate synchronized nonverbal behavior. Kopp and colleagues [25,25] developed a comprehensive model for gesture animation based on research in psycholinguistics and motor control theory. Here, a knowledge base, similar to de Ruiter's gestuary [11], holds gesture templates which consist of hierarchies of constraints on static and dynamic features of the stroke phase. Gesture production instantiates templates and feeds them into a motor planner for execution. Preparation, retraction and co-articulation effects are automatically appended. The model supports sophisticated arm trajectories including velocity profiles. The system also supports speech parameterization through SABLE [27]. Recently, Cassell, Kopp and colleagues brought together the best from the aforementioned systems in *NUMACK* [28], a system capable of synthesizing in real-time co-verbal context-sensitive iconic gestures without relying on a library of predefined gestures. Though the gesture-speech production process is not the focus of the chapter, the underlying gesticulation animation model in these systems shares several aspects with the model presented in section 5.4, namely: the requisites are based on psycholinguistics research and similar static and dynamic features are explored.

5.3.3 Interface Languages

Controlling and integrating gesticulation expression with other modalities is usually solved through markup languages [29]. The idea is that the gesticulation production process communicates the gesticulation plan, created from the communicative intent, to the gesticulation execution process, which animates it, through this language. The language, thus, supports a convenient clear-cut separation between these processes. Presently, no such standard language exists. The community has acknowledged this and has begun to address it. A promising effort is the SAIBA framework [30] which brings together several research groups. Unfortunately, this standard is still in its infancy and, therefore, the model presented in section 5.4 requires, for the time being, yet another control language – *Expression Markup Language (EML)*. This language is particularly influenced by: VHML [31], SMIL [32] and MURML [33]. Regarding Virtual Human Markup Language (VHML), this work reuses the notion of organizing control according to modality-specific modules. Regarding Synchronized Multimedia Integration Language (SMIL), which is oriented towards audiovisual interactive presentations, this work uses a similar modality synchronization mechanism. Regarding Multimodal Utterance Representation Markup Language (MURML), this work defines a similar notation for gesture specification and synchronization with co-verbal speech. Finally, in contrast to high-level languages such as GESTYLE [34] which tries to capture the individual's expression style and APML [35] which represents, among others, communicative intent, emotions, interaction and cultural aspects, the proposed language focuses on speech synthesis and low-level body control such as gesticulation animation as sequences of constraints on static and dynamic features.

5.4 A Model for Gesticulation Expression

This section describes a gesticulation expression model for virtual humans which supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientations and positions) and dynamic features; (b) multimodal synchronization between gesticulation and speech; (c) automatic reproduction of annotated gesticulation according to GestuRA, a transcription algorithm; (d) expression control through an abstract integrated synchronized language. The model builds on top of a virtual human architecture, which provides keyframe and procedural animation mechanisms, and integrates with other expression modalities including speech, Fig. 5.2.

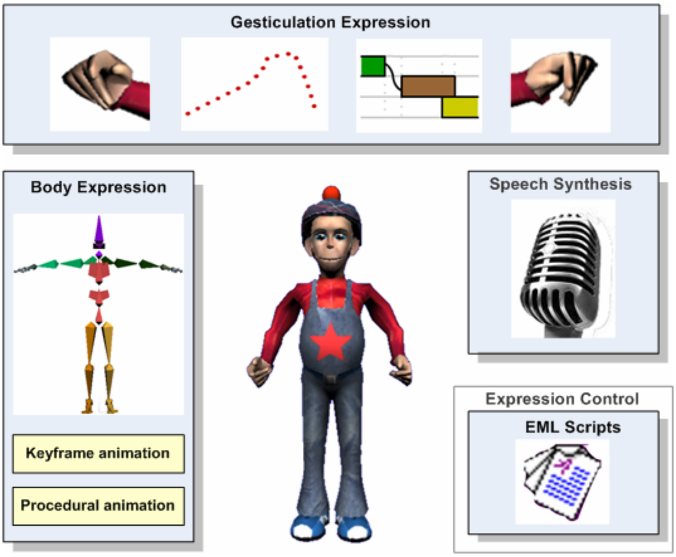


Fig. 5.2. Gesticulation expression model overview

The remainder of this section is organized as follows: subsection 5.4.1 describes the virtual human architecture; subsection 5.4.2 describes speech synthesis and integration with gesticulation; subsection 5.4.3 describes the feature-based gesticulation model itself; subsection 5.4.4 describes multimodal expression control using a markup scripting language and subsection 5.4.5 describes two evaluation studies.

5.4.1 Virtual Humans

The virtual human is structured according to a three-layer architecture [36,37]. The *geometry layer* defines a 54-bone human-based skeleton. The *animation layer* defines keyframe and procedural animation mechanisms. The *behavior layer* defines speech and gesticulation expression and supports a language for integrated synchronized multimodal expression.

5.4.1.1 Keyframe Animation

Keyframe animation animates the virtual human according to predefined sequences of poses usually designed by human artists. The model generates in-between poses, supports several animation combination mechanisms but, ultimately, is not independent of the human creator and, thus, is not very flexible. Still, keyframe animation is useful for gesticulation expression in the following situations: (a) animation of complex gesticulation which is cumbersome to model through features; (b) animation of gesticulation which involves body parts other than arms and hands. Keyframe animation revolves around *animation players* which animate subsets of the skeleton's bones according to specific animation mechanisms. Several players can be active at the same time and thus, as they may compete for the same bones, an arbitration mechanism based on priorities is defined. Supported animation mechanisms include: (a) *weighted combined animation*, where the resulting animation is the "weighted average" of animations placed on several weighted layers; (b) *body group animation*, where disjoint sets of skeleton's bones – body groups – execute independent animations; (c) *pose animation*, which applies stances to bones, supports combination between two stances and provides a parameter to control interpolation between them.

5.4.1.2 Procedural Animation

Procedural animation consists of animating the virtual humans by controlling the limbs' extremities. Procedural animation is at the core of flexible gesticulation expression as it provides the means to position and orient the hands arbitrarily in space according to specific motion profiles. Notice this flexibility isn't possible using keyframe animation. Procedural animation is based on robotics techniques [38]. In the geometry layer, six revolute joint robotic manipulators are integrated with the skeleton to control the limbs and joint limits are defined according to anthropometry data [39]. In the animation layer, three inverse kinematics and one inverse velocity primitives are defined: (1) *joint interpolation*, which animates the manipulator's target through interpolation in the joint space; (2) *function based interpolation*, which animates the target according to a transformation defined, at each instant, by a mathematical function; (3) *frame interpolation*, which animates the target according to interpolation between the current frame and the intended frame; (4) *Jacobian-based animation*, which applies inverse velocity algorithms to animate the target according to intended Cartesian and angular velocities.

5.4.2 Voice Synthesis

Voice synthesis is based on the Festival [40] text-to-speech system. Festival features facilitate integration with gesticulation as they include: (a) a simple Scheme programming interface; (b) server/client interaction through sockets thus, supporting clients in other programming languages; (c) access to synthesized utterance structure (words, phonemes, times, etc.), which synchronizes with gesticulation phases, and the ability to save this data in files; (d) incremental real-time synthesis, thus, allowing the virtual human to schedule gesticulation while speech is being synthesized; (e) limited support for SABLE [27] which allows definition of speech emphasis, prosodic breaks, velocity, pitch, text volume configuration, among others.

Festival integration with the virtual human involves four aspects, Fig. 5.3: (1) the notion of speech; (2) an extension to Festival's voice synthesis pipeline; (3) a communication protocol; (4) a new behavior layer API for speech control. A speech is modeled as a set of files including: (a) utterance structure, i.e., phonemes, words and times; (b) utterance waveforms; (c) a configuration file with information about all files. Using Festival's programming interface, the voice synthesis pipeline is extended, after natural language and signal processing, with the following steps: after each utterance has been synthesized, its structure and waveform are saved and the virtual human is informed that an utterance is ready to play; after all utterances have been synthesized, the speech file is saved and the virtual human is informed about speech synthesis completion. The communication protocol is characterized as follows: (a) supports voice synthesis primitives; (b) supports incremental utterance conclusion communication; (c) supports communication of speech synthesis conclusion. At the virtual

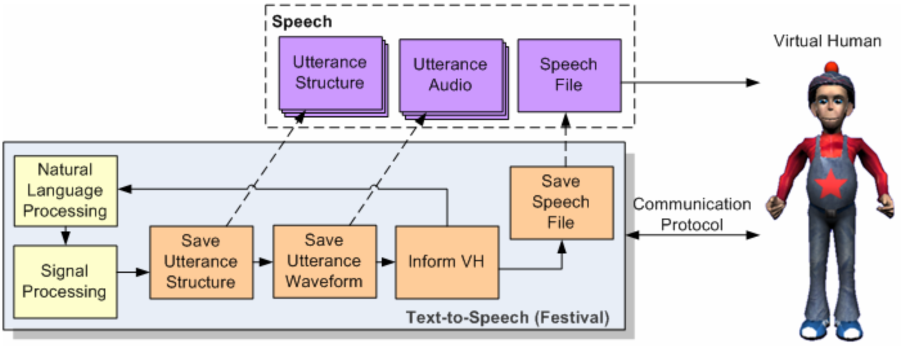


Fig. 5.3. Speech-synthesis integration with Festival

human side, the behavior layer was extended to support two voice primitives: (1) *synchronous text-to-speech*, which initiates voice synthesis with real-time feedback as utterances are synthesized; (2) *preprocess text*, which synthesizes speech and saves it in a persistent format for posterior playback. Both primitives support SABLE.

5.4.3 Gesticulation Expression

The gesticulation expression model controls arms and hands and relies on keyframe and procedural animation. Precisely, limb manipulators control the arms, hands' position and orientation while pose animation players control the hands' shape. The model is feature-based, i.e., gesticulation form is modeled as a sequence in time of constraints on static and dynamic features. Features are described on subsection 5.4.3.1. Motion modifiers influence the interpretation of otherwise neutral gesticulation. Modifiers are described on subsection 5.4.3.2. The model supports multimodal synchronization, in particular, between speech and gesture. Synchronization is described on subsection 5.4.3.3. Finally, the model supports automatic reproduction of annotated gesticulation according to GestuRA, a gesture transcription algorithm. GestuRA and its integration with the model are described on subsection 5.4.3.4.

5.4.3.1 Features

Gesticulation is modeled as a sequence in time of constraints on static and dynamic features. Static features are represented in *gesticulation keyframes* and include: hand shape, position, orientation palm axis, orientation angle, and handedness. Dynamic features define motion profiles for keyframe interpolation.

Regarding static features, the *hand shape* feature can assume any Portuguese Sign Language hand shape [4]. Furthermore, any two shapes can be combined and a parameter is provided to define how much each contributes. Implementation relies on pose player ability to combine stances and on a library of stances for Portuguese Sign Language shapes. The *position* feature is defined in Cartesian coordinates in three-dimensional space. Both world and speaker references can be used. Hand shape orientation is defined by two features: *orientation palm axis*, which defines the palm's normal; and *orientation angle* which defines a left handed angle about the normal. Implementation relies on inverse kinematics primitives. The *handedness* feature defines whether the gesticulation keyframe applies to the left, right or both hands. In the last case, remaining features apply to the speaker's dominant hand and symmetrical values apply to the non-dominant hand. Symmetry is intuitively understood as the gesticulation which would result if a mirror stood on the sagittal plane.

Regarding dynamic features, the model supports (keyframe) interpolation through parametric cubic curves, which can represent any kind of velocity profile, such as deceleration near the target position and overshooting effects which we see in humans [41]. Currently, the model supports Bézier and Hermite cubic curves, as well as piecewise combinations thereof. Furthermore, interpolators can be structured into hierarchies thus, leading to sophisticated motion profiles. Moreover, either Cartesian or joint angle velocity can be used. Implementation of interpolation in Cartesian and joint angle space relies, respectively, on the frame interpolation and joint interpolation procedural animation control primitives.

5.4.3.2 Modifiers

Several researchers have explored *motion modifiers* which add emotive qualities to existent motion data. Signal-processing techniques [42,43,44] were used to extract information from motion data which is used to generate emotional variations of neutral motion. Rose and colleagues [45] generate new motion with a certain mood or emotion from motion data interpolation based on radial functions and low order polynomials. Chi and colleagues [46] propose a system which adds expressiveness to existent motion data based on the effort and shape parameters of a dance movement observation technique called Laban Movement Analysis. Finally, Hartmann [47] draws from psychology six parameters for gesture modification: *overall activation*, which refers to the quantity of movement during a conversational turn; *spatial extent*, which refers to the amplitude of movement; *temporal extent*, which refers to the duration of movements; *fluidity*, which refers to smoothness and continuity of movement; *power*, which refers to how strong or weak the movement appears; and *repetition*, which refers to rhythmic repeats of specific movement.

The effect of these modifiers can be simulated resorting to the static and dynamic features described above. However, in digital worlds, motion modifiers need not be limited to the body. Thus, inspiring in the arts, we've explored a different set of modifiers which rely on properties of the surrounding environment [48] – such as camera, lights and music – and the screen [49] – such as the virtual human pixels themselves – to convey emotional interpretations to virtual human movement. These modifiers are, however, detailed elsewhere [48,49].

5.4.3.3 Synchronization

Sub-second synchronization of gesture phases with speech relies on a control markup language – Expression Markup Language (EML) – which supports phoneme-level synchronization. The language integrates with SABLE [27] and thus, supports synchronization with speech properties such as intonation contour. Similarly to SMIL [32], modality execution time can be set to absolute or modality relative values. Furthermore, named timestamps can be associated with text to be synthesized. The following events can be associated with named timestamps: (a) start of a word; (b) end of a word; (c) start of a phoneme. EML is detailed on subsection 5.4.4.

As synchronization between speech and gesture is conveniently described at the gesture phase level, the model supports explicit *gesticulation phase keyframes*. The phase keyframe extends regular keyframes as follows: (a) a duration feature is added which defines total phase time; (b) sequences of constraints can now be associated to shape, position and orientation features; (c) constraints within a sequence can be set to start at absolute time offsets relative to phase start time or at percentages of the total phase duration. However, phase keyframes do not add expressiveness to the model in the sense that gesticulation described with phase keyframes could be converted into an equivalent sequence of regular keyframes.

5.4.3.4 Automatic Reproduction of Gesticulation Annotations

The gesticulation model supports automatic reproduction of *Gesture Recording Algorithm (GestuRA) annotations*. GestuRA, based on [2] and [50], is a linguistically motivated iterative algorithm for gesticulation *form* and *meaning* transcription. The former refers to the kinesthetic properties, whereas the latter to the interpretation of the gesture. GestuRA is structured into seven passes, Fig. 5.4. First, speech is transcribed from the video-speech record. Second, text is organized into utterances. Third, utterances are classified according to discourse levels – narrative, metanarrative and paranarrative [1]. Fourth, gesticulation is filtered ignoring remaining gestures (such as emblems, for instance). Fifth, gesticulation phases are annotated. Sixth, gesticulation form is formally annotated. Finally, seventh, gesticulation is classified according to its dimensions and its meaning analyzed.

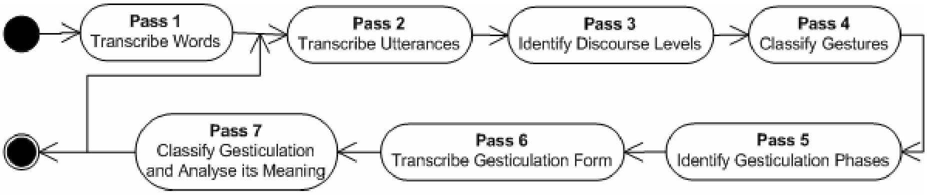

Pass 1
Transcribe Words
Pass 2
Transcribe Utterances
Pass 3
Identify Discourse Levels
Pass 4
Classify Gestures
Pass 5
Identify Gesticulation Phases
Pass 6
Transcribe Gesticulation Form
Pass 7
Classify Gesticulation and Analyse its Meaning

Fig. 5.4. Overview of the Gesture Recording Algorithm (GestuRA)

GestuRA integration with the gesticulation model is achieved through *Anvil* [51], a generic multimodal annotation tool. In concrete, implementing GestuRA in Anvil benefits from its capability of exporting annotations to a XML format. This format can, then, be converted into EML for immediate execution in virtual humans, Fig. 5.5.

Automatic reproduction from GestuRA is valuable for various reasons. First, reproduction from transcribed annotations is flexible. Usually annotation algorithms are used to build databases of human gestures. Thus, all gesture details need to be formalized and are, usually, classified according to form and meaning. Therefore, reproduction from such an annotation can selectively choose which information to use according to context. For instance, if we want to provide a virtual human with a certain style, we could disregard form annotation and simply reproduce gestures from the annotated meaning but, using a stylized form. This flexibility contrasts with reproduction from automatic gesture recognition algorithms [52,53], which accurately recognize form but, are still limited with respect to meaning interpretation. Automatic reproduction is also useful

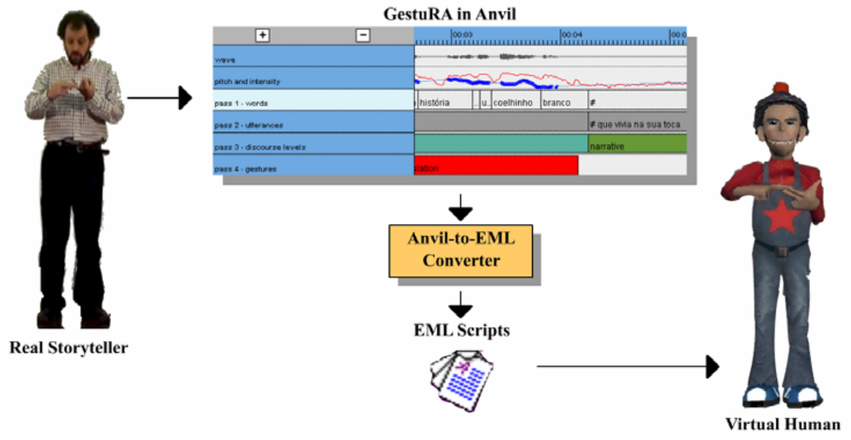


Fig. 5.5. Gesticulation model integration with GestuRA

to test transcription accuracy and, furthermore, constitutes an important evaluation tool for the gesticulation expression model. As speech and gesture production from communicative intent is not simulated, an alternative to evaluating the model is to compare it to actual real-life videos.

5.4.4 Multimodal Expression Control

This work proposes a markup, integrated and synchronized language – *Expression Markup Language (EML)* – which serves as a control interface for the body. The language can be used in two ways, Fig. 5.6: (1) as an *interface for a mind* which needs to express, in real-time, synchronously and multimodaly through the body; (2) as a *script* which describes a story, written by an author, where the virtual human expresses multimodaly. In the first case, the mind communicates to the body in real-time, through a socket or API, a set of EML clauses which are immediately executed. The gesticulation production process is meant to integrate with the execution process in this way (see section 5.3). In the second case, the script defines a sequence of

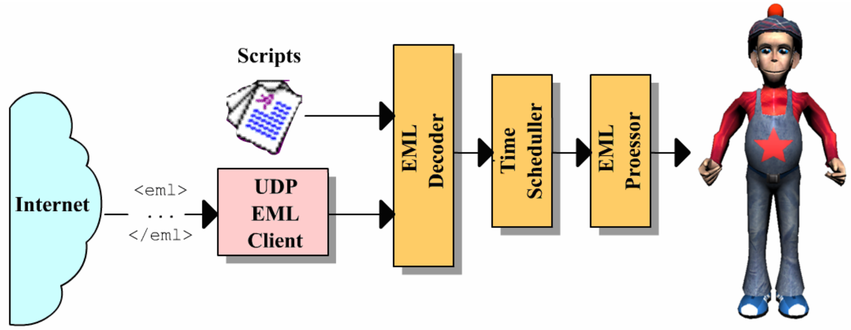


Fig. 5.6. EML integration with the virtual human

clauses, temporally ordered, which defines a story which can be played later by different virtual humans. Regarding specification, EML is a markup language structured into modules: (1) *core*, defines the main elements; (2) *time and synchronization,* defines multimodal synchronization and is characterized as follows: (a) supports execution time definition relative to other clauses; (b) supports execution time definition relative to word or phoneme in vocal expression clauses; (c) supports loops; (d) supports parallel and sequential execution. This module is based on W3C's SMIL 2.0 specification [32]; (3) *body,* controls both keyframe and procedural animation; (4) *voice,* controls speech synthesis; (5) *gesture*, controls gesticulation expression.

5.4.5 Evaluation

Two studies were conducted to assess the model's expressiveness. In both cases, the idea consisted of comparing the narration of the Portuguese traditional story "The White Rabbit" by a human storyteller with a version by a virtual storyteller. The first study, conducted in the scope of the "Papous" project at Inesc-ID, aimed at evaluating several expression modalities while the second focused only on gesticulation.

5.4.5.1 First Study

The first study was conducted in the scope of the "Papous" project at Inesc-ID[1] and aimed at comparing a human storyteller with a virtual storyteller with respect to story comprehension, emotion expression, credibility and subject satisfaction for each of gesticulation, facial and vocal expression. This document will focus only on the gesticulation expression results. The human storyteller was a non-professional actor which was simply asked to tell the story in an expressive way without imposing any requirements on gesticulation expression. Regarding the virtual storyteller, the voice consisted of modulated synthesized speech audio records. Facial expression, including proper lip-synch and emotion expression, was generated from a pseudo-muscular model [54]. Gesticulation expression was based on a GestuRA transcription of the human storyteller video, lasting 7 minutes and 30 seconds. In total, 286 gestures were transcribed of which 95% were automatically reproduced through feature-based gesticulation expression and 5% through keyframe animation.

Regarding structure, the subject begins by visualizing the story video and, then, answers to a questionnaire. Each subject is presented with one of four video versions: (1) *CRVR*, which uses the human narrator with real voice; (2) *CRVS*, which uses the human narrator with synthetic voice; (3) *CSVR*, which uses the virtual narrator with real voice; (4) *CSVS*, which uses the virtual narrator with synthetic voice. The questionnaire consists of twelve classification questions where the subject is asked to classify, from 1 (totally disagree) to 7 (totally agree), whether each modality helps understand the story, expresses emotions properly, is believable and is to his liking.

The study was presented to 108 students at IST-Technical University of Lisbon. Average age was 21 years and 89% were males. Most students were enrolled in technology-related courses. Each video version was presented to 27 students.

Gesticulation expression results are summarized in Table 5.1. Comparing real and synthetic gestures classifications, it is clear that real gestures got the best classification for

[1] Papous: Papous project at Inesc-ID. Ref.: POSI / SRI / 41071 / 2001.

Table 5.1. Summary of results for gesture classification questions in the first study

	CRVR	CSVR	CRVS	CSVS
Did gestures help understand the story?				
Negative (%)	7.4	22.2	11.1	14.8
Neutral (%)	7.4	3.7	7.4	7.4
Positive (%)	85.2	74.1	81.5	77.8
Did gestures express the story's emotions?				
Negative (%)	7.4	29.6	3.7	18.5
Neutral (%)	11.1	7.4	3.7	14.8
Positive (%)	81.5	63	92.6	66.7
Were gestures believable?				
Negative (%)	11.1	37	11.1	18.5
Neutral (%)	14.8	3.7	11.1	25.9
Positive (%)	74.1	59.3	77.8	55.6
Did you like the gestures?				
Negative (%)	11.1	29.6	7.4	11.1
Neutral (%)	22.2	11.1	3.7	22.2
Positive (%)	66.7	59.3	88.9	66.7

every question. Nevertheless, synthetic gestures were positively classified by the majority of subjects with respect to contribution to story comprehension, emotion expression, believability and liking (positive answers above 50% for all questions).

From these results it is possible to conclude that synthetic gestures contribute to story comprehension, emotion expression and believability. Still, synthetic gestures do not capture all the subtleties of its real counterpart as these are better classified in general. Furthermore, this study had some limitations. Firstly, subjects were asked to evaluate gestures explicitly when it is known that gesture interpretation is essentially unconscious [1,2]. Secondly, subject to multiple interpretations, the notion of "believability" is hard to define thus, results related to the question "Gestures were believable" should be interpreted with caution.

5.4.5.2 Second Study

To further assess the model's expressiveness and to correct some of the flaws in the previous study, a second study was conducted. In this study, first, subjects are told that the evaluation is about virtual storytelling and "gesticulation expression" is never mentioned throughout. Second, synthetic gestures are indirectly evaluated through story interpretation questions. Third, each subject sees the story alternatively narrated by the human or virtual storyteller thus, allowing for direct comparison. Finally, as the study focuses on gesticulation expression, the real voice is used for both storytellers and three variations of the virtual storyteller are defined: (1) *ST*, which uses both feature-based and keyframe gesticulation; (2) *SF*, which uses only feature-based gesticulation; (3) *SN*, which uses no gesticulation.

The evaluation is structured into three parts. In part 1 – *profile* – the subject profile is assessed. In part 2 – *story interpretation* – the whole story is presented. To facilitate remembering, the story is divided into 8 segments of 30 seconds each. Segments are narrated by either the human storyteller or one of the three kinds of virtual storytellers randomly selected at the start. In concrete, the third and sixth segments are narrated by a subject selected storyteller, while the rest is arbitrarily narrated either by the human or virtual storyteller provided that in the end each narrates an equal number of segments. After each segment, multiple choice interpretation questions are posed. In total 32 questions were formulated. Importantly, a subset, named the *highly bodily expressive (HBE)* questions, focuses on information specially marked in gestures, i.e., information which is either redundantly or non-redundantly conveyed through complex gestures like iconics or metaphorics. Finally, in part 3 – *story appreciation* – the subject is asked to choose the preferred storyteller and to describe the best and worst feature of each storyteller.

The study was presented to 39 subjects, 90% of which were male, with average age of 23 years and most had college-level education. The study was fully automated in software and average evaluation time was about 20 minutes. Distribution of virtual storyteller kinds across subjects was: 46% for ST; 31% for SF; 23% for SN. Subject recruitment included personal contact mainly at both campuses of IST-Technical University of Lisbon and distribution of the software through the Web.

Regarding story interpretation results, if we define *diff* to be the difference between the percentage of correct answers following the human storyteller and the percentage of correct answers following the virtual storyteller, then *diff* was: for ST, 4.69%; for SF, -0.68%; for SN, -1.62%. However, if we consider only HBE questions, than distribution is as follows: for ST, 4.75%; for SF, 0.00%; for SN, 9.19%. Regarding subject storyteller selection on the third and sixth segments, the human storyteller was selected about 75% of the time (for ST, 75.00%; for SF, 83.30%; for SN, 72.22%). Regarding subject storyteller preference, the human storyteller was preferred about 90% of the time (for ST, 88.89%; for SF, 83.33%; for SN, 100.00%). Finally, some of the worst aspects mentioned for the virtual storyteller were "body expression limited to arms", "static/rigid", "artificial" and "low expressivity". These relate to the best aspects mentioned for the human storyteller, namely "varied postures", "energetic/enthusiastic", "natural" and "high expressivity".

As can be seen by these results, the human storyteller fares better than the virtual storyteller. Interpretation with the human storyteller is better, though not that much (*diff* of 4.69% for ST). Furthermore, when given a choice, subjects almost always choose the human storyteller. Analyzing the best and worst aspects selected for each storyteller might give insight into this issue. Surprisingly, if all questions are considered, *diff* actually reduces for SN when compared to ST (-1.63% over 4.69%). The fact that the human storyteller's voice and face were highly expressive and gestures were mostly redundant might help explain this. However, if only HBE questions are considered, *diff* considerably increases for the SN case (from 4.75% to 9.19%).

Furthermore, for the SN case, the human storyteller was preferred 100% of the times. This confirms that gesticulation affects interpretation. Finally, comparing ST with SF, *diff* for all questions reduces for the latter case (from 4.69% to -0.68%). This suggests that the lack of feature-based gesticulation support for the small fraction of highly complex gestures does not impede effective interpretation.

5.5 Discussion and Future Work

This chapter overviews the challenge of building a virtual human computational model of gesticulation expression. First, a virtual human architecture is required with appropriate control mechanisms to support gesticulation animation. Second, the gesticulation execution problem, which refers to converting a gesticulation plan into an animation plan, must be addressed. Finally, the speech-gesticulation problem, which refers to converting communicative intent into verbal and gesticulation plans, should be addressed.

The chapter also proposes a gesticulation expression model which supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientation palm axis, orientation angle and handedness) and dynamic features; (b) multimodal synchronization between gesticulation and speech; (c) automatic reproduction of GestuRA annotations; (d) expression control through the abstract integrated synchronized Expression Markup Language. The model builds on top of a layered virtual human architecture which supports keyframe and procedural animation. Finally, two studies were conducted to evaluate the model in a storytelling context. In both cases, we compare the expression of a human with a virtual storyteller. Results indicate that synthetic gestures contributed to story comprehension, emotion expression and believability. Furthermore, synthetic gestures fared well against real gestures. However, the fact that the human storyteller was consistently preferred hints that there is still room for improvement.

Therefore, regarding future work, first, gesticulation needs to go beyond arms and hands and explore other body parts. Posture shifts, which relate to discourse structure [55], could be explored. Second, some features' implementation restrict expressiveness. For instance, nothing guarantees that Portuguese Sign Language hand shapes and non-spline parametric curves (such as Bézier and Hermite) and combinations thereof suffice to express, respectively, all shapes and motion profiles. Furthermore, lack of elbow control in the upper limb manipulator limits naturalness [38]. Third, preparation and retraction motion, as well as co-articulation effects, could be automatically generated. Finally, a more anatomically correct hand model with appropriate constraints (subsection 5.3.1) would lead to more realistic gesticulation simulation.

At a more global level, the next step is to tackle the gesticulation production problem. Altogether, the model seems ready to support speech and gesticulation production models (subsection 5.2.4). Regarding de Ruiter's model, the gestuary can mostly be implemented through feature-based and keyframe gesticulation; signal passing synchronization is straightforwardly supported. Krauss' model which is feature-based is also compatible with the model but, cross-modal priming is not supported. The language effect on gesture in Kita and Özyürek's model occurs early in the production process and, ultimately, materializes into features which the model supports.

McNeill's growth point model doesn't detail morphology generation. However, if the dialectic ultimately materializes into features and synchronization can be described with a finite number of synchronization points, then the model is likely to support it.

Acknowledgments

This research was partially supported by the Papous project at Inesc-ID (Ref.: POSI / SRI / 41071 / 2001).

References

1. McNeill, D.: Hand and Mind: What gestures reveal about thought. University of Chicago Press (1992)
2. McNeill, D.: Gesture and Thought. University of Chicago Press (2005)
3. Kendon, A.: How gestures can become like words. In: Poyatos, F. (ed.) Cross-cultural perspectives in nonverbal communication, Hogrefe, pp. 131–141 (1988)
4. Secretariado Nacional para a Reabilitação e Integração das Pessoas com Deficiência: Gestuário – Língua Gestual Portuguesa, 5th edn. (1991)
5. Kendon, A.: Some relationships between body motion and speech. In: Siegman, A., Pope, B. (eds.) Studies in dyadic communication, pp. 177–210. Pergamon Press, New York (1972)
6. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. In: Key, M. (ed.) The Relationship of Verbal and Nonverbal Communication, pp. 207–227. Mouton and Co. (1980)
7. Nobe, S.: Where do most spontaneous representational gestures actually occur with respect to speech? In: McNeill, D. (ed.) Language and Gesture, pp. 186–198. Cambridge University Press, Cambridge (2000)
8. Kendon, A.: Sign languages of Aboriginal Australia: Cultural, semiotic and communicative perspectives. Cambridge University Press, Cambridge (1988)
9. Kita, S.: The temporal relationship between gesture and speech: A study of Japanese-English bilingual. MhD thesis, Department of Psychology, University of Chicago (1990)
10. Levelt, W.: Speaking. MIT Press, Cambridge (1989)
11. de Ruiter, J.: The production of gesture and speech. In: McNeill, D. (ed.) Language and gesture, pp. 284–311. Cambridge University Press, Cambridge (2000)
12. Krauss, M., Chen, Y., Gottesman, R.: Lexical gestures and lexical access: A process model. In: McNeill, D. (ed.) Language and gesture, pp. 261–283. Cambridge University Press, Cambridge (2000)
13. Kita, S., Özyürek, A.: What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking in Journal of Memory and Language 48, 16–32 (2003)
14. Akenine-Möller, T., Haines, E.: Real-time Rendering, 2nd edn. A K Peters (2002)
15. Cavazza, M., Earnshaw, R., Magnenat-Thalmann, N., Thalmann, D.: Motion Control of Virtual Humans in IEEE Computer Graphics and Applications 18(5), 24–31 (1998)
16. Thompson, D., Buford, W., Myers, L., Giurintano, D., Brewer III, J.: A Hand Biomechanics Workstation in Computer Graphics 22(4), 335–343 (1988)
17. Wagner, C.: The pianist's hand: Anthropometry and biomechanics in Ergonomics 31(1), 97–131 (1988)

18. Moccozet, L., Magnenat-Thalmann, N.: Dirichlet Free-Form Deformations and their Application to Hand Simulation. In: Proc. Computer Animation 1997, pp. 93–102 (1997)
19. Sibille, L., Teschner, M., Srivastava, S., Latombe, J.: Interactive Simulation of the Human Hand. In: CARS 2002, pp. 7–12 (2002)
20. Albrecht, I., Haber, J.H., Siedel, H.: Construction and Animation of Anatomically Based Human Hand Models. In: SIGGRAPH 2003, pp. 98–109 (2003)
21. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agent. In: Proc. of SIGGRAPH 1994, pp. 413–420 (1994)
22. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in Conversational Interfaces: Rea. In: Proc. of the CHI 1999 Conference, Pittsburgh, PA, pp. 520–527 (1999)
23. Cassell, J., Stone, M.: Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. In: Proc. of the AAAI 1999 Fall Symposium on Psychological Models of Communication in Collaborative Systems, North Falmouth, MA, pp. 34–42 (1999)
24. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: Proc. of SIGGRAPH 2001, pp. 477–486 (2001)
25. Kopp, S., Wachsmuth, I.: A knowledge-based approach for lifelike gesture animation. In: Proc. of the 14th European Conf. on Artificial Intelligence. IOS Press, Amsterdam (2000)
26. Wachsmuth, I., Kopp, S.: Lifelike Gesture Synthesis and Timing for Conversational Agents. In: Wachsmuth, I., Sowa, T. (eds.) GW 2001. LNCS (LNAI), vol. 2298, pp. 120–133. Springer, Heidelberg (2002)
27. SABLE: A Synthesis Markup Language (v. 1.0), http://www.bell-labs.com/project/tts/sable.html
28. Kopp, S., Tepper, P., Cassell, J.: Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output. In: Proc. of the International Conference on Multimodal Interfaces (ICMI 2004), pp. 97–104. ACM Press, New York (2004)
29. Arafa, Y., Kamyab, K., Mamdani, E.: Character Animation Scripting Languages: A Comparison. In: Proc. of the International Conference on Autonomous Agents 2003, Melbourne, Australia (2003)
30. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., Vilhjálmsson, H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In: Proc. of Intelligent Virtual Agents 2006, pp. 205–217 (2006)
31. VHML: VHML – Virtual Human Markup Language, `http://www.vhml.org/`
32. SMIL: SMIL - Synchronized Multimedia http://www.w3.org/AudioVideo/
33. Kranstedt, A., Kopp, S., Wachsmuth, I.: MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In: Falcone, R., Barber, S., Korba, L., Singh, M.P. (eds.) AAMAS 2002. LNCS (LNAI), vol. 2631. Springer, Heidelberg (2003)
34. Ruttkay, Z., Noot, H.: Variations in Gesturing and Speech by GESTYLE in International Journal of Human-Computer Studies. Special Issue on Subtle Expressivity for Characters and Robots 62(2), 211–229 (2005)
35. de Carolis, B., Pelachaud, C., Poggi, I., Steedman, M.: APML, a Mark-up Language for Believable Behavior Generation. In: Prendinger, H. (ed.) Life-like Characters. Tools, Affective Functions and Applications. Springer, Heidelberg (2004)
36. Blumberg, B., Galyean, T.: Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments. In: Proc. of SIGGRAPH 1995, vol. 30(3), pp. 47–54 (1995)

37. Perlin, K., Goldberg, A.: Improv: A System for Scripting Interactive Actors in Virtual Worlds. In: Proc. of SIGGRAPH 1996, pp. 205–216 (1996)
38. Tolani, D., Goswani, A., Badler, N.: Real-time inverse kinematics techniques for anthropomorphic limbs in Graphics Models 62, 353–338 (2000)
39. NASA Man-Systems Integration Manual (NASA-STD-3000)
40. Festival: The Festival Speech Synthesis Systems, http://www.cstr.ed.ac.uk/projects/festival/
41. Mark, L.: Control of Human Movement. Human Kinetics Publishers (1993)
42. Unuma, M., Anjou, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: Proc. of SIGGRAPH 1995, pp. 91–96 (1995)
43. Brudelin, A., Williams, L.: Motion signal processing. In: Proc. of SIGGRAPH 1995, pp. 97–104 (1995)
44. Amaya, K., Bruderlin, A., Calvert, T.: Emotion from motion. In: Proc. Graphics Interface 1996, pp. 222–229 (1996)
45. Rose, C., Bodenheimer, B., Cohen, M.: Verbs and Adverbs: Multidimensional Motion Interpolation. IEEE Computer Graphics and Applications 18(5), 32–40 (1998)
46. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: Proc. of SIGGRAPH 2000, pp. 173–182 (2000)
47. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents in Gesture Workshop. Springer, Heidelberg (2005)
48. de Melo, C., Paiva, A.: Environment Expression: Expressing Emotions through Cameras, Lights and Music. In: Proc. of Affective Computing Intelligent Interaction 2005, pp. 715–722 (2005)
49. de Melo, C., Paiva, A.: Expression of Emotions in Virtual Humans using Lights, Shadows, Composition and Filters. In: Proc. of Affective Computing Intelligent Interaction 2007, pp. 546–557 (2007)
50. Gut, U., Looks, K., Thies, A., Trippel, T., Gibbon, D.: CoGest – Conversational Gesture Transcription System, Technical Report. University of Bielefeld (1993)
51. Kipp, M.: ANVIL – A Generic Annotation Tool for Multimodal Dialogue. In: Proc. of the 7th European Conference on Speech Communication and Technology, pp. 1367–1370 (2001)
52. Pavlovic, V., Sharma, R., Huang, T.: Visual Interpretation of hand gestures for human computer interaction. A review in IEEE Trans. Pattern Analysis Machine Intelligence 19, 677–695 (1997)
53. Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding 73, 82–98 (1999)
54. Raimundo, G.: Real-time Facial Expression and Animation. MSc thesis, Department of Information Systems and Computer Engineering, IST-Technical University of Lisbon (2007)
55. Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C.: Annotating and Generating Posture from Discourse Structure in Embodied Conversational Agents. In: Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, Autonomous Agents 2001 (2001)

6

MIKI: A Case Study of an Intelligent Kiosk and Its Usability

Lee McCauley, Sidney D'Mello, Loel Kim, and Melanie Polkosky

The University of Memphis
Memphis, TN 38152, USA
{mccauley,sdmello,lkim}@memphis.edu,
mpolkosky@cmamemphis.com

Abstract. MIKI is a three-dimensional directory assistance-type digital persona displayed on a prominently-positioned 50 inch plasma unit housed at the FedEx Institute of Technology at the University of Memphis. MIKI, which stands for Memphis Intelligent Kiosk Initiative, guides students, faculty and visitors through the Institute's maze of classrooms, labs, lecture halls and offices through graphically-rich, multidimensional, interactive, touch and voice sensitive digital content. MIKI differs from other intelligent kiosk systems by its advanced natural language understanding capabilities that provide it with the ability to answer informal verbal queries without the need for rigorous phraseology. This chapter first describes, in general, the design and implementation of the Intelligent Kiosk. We then describe a usability study conducted to evaluate the functionality of the system. While the usability testing exemplified good interface design in a number of areas, the complexity of multiple modalities—including animated graphics, speech technology and an avatar greeter—complicated usability testing, leaving developers seeking improved instruments. In particular, factors such as gender and technical background of the user seemed to change the way that various kiosk tasks were perceived, deficiencies were observed in speech interaction as well as the location information in a 3D animated map.

6.1 Introduction

As we find ourselves at the height of the information age the need for user-friendly, naturalistic, and intuitive information systems becomes paramount. Stephandis et al. have predicted that public information systems, terminals, and information appliances will be increasingly used in a variety of domains [1]. Of particular interest in the field of intelligent information systems and virtual agents are information kiosks. These are a special variant of information appliances that are usually deployed in public locations such as transportation hubs, malls, businesses, etc. In addition to the basic issues that accompany the design of typical information systems such as information retrieval, multi-modal communication, and interface design these systems pose some novel and interesting concerns. Cassell et al., point out that kiosk systems differ from traditional systems in that they should stand out so that they are noticed by visitors, their functions should be self-evident, no user training should be required, and they should be able to recover from user errors [2].

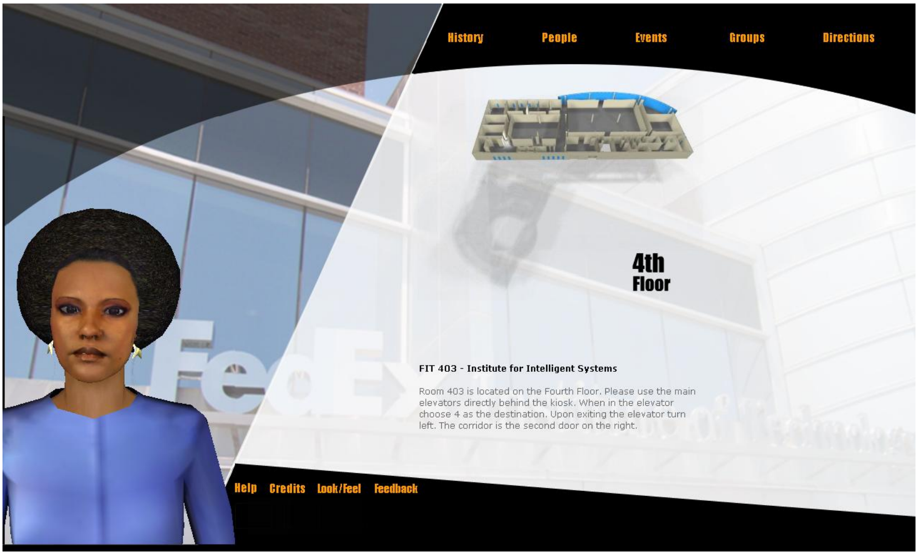
One of the information kiosks that demonstrated a significant improvement over earlier systems is the MINNELLI system [3]. MINNELLI facilitates interactions with bank customers primarily by the use of short animated cartoons that present

N. Magnenat-Thalmann et al. (Eds.): New Adv. Virt. Hum.: Arti. Intel., SCI 140, pp. 153–176, 2008.
springerlink.com

information on bank services. However, the MINNELLI system requires basic user training which reduces its applicability in most public sites. Another successful kiosk with a broader scope than the MINELLI system is the MACK system [2]. MACK is an embodied conversational kiosk that provides information on residents and directions to locations at a research site. It integrates multiple input sources that include speech, gesture, and pressure. The system also exhibits a degree of spatial intelligence by utilizing its awareness of its location and the layout of the building to reference physical locations when it provides directions [4]. The August spoken dialog system is also kiosk based and helps users find their way around Stockholm, Sweden using an on-screen street map. August is designed to elicit a conversation from the user and facilitates the study of such interactions [5, 6].

This chapter describes the design, implementation, and usability studies conducted on one such intelligent kiosk system called MIKI: The Memphis Intelligent Kiosk Initiative. The system is deployed on a plasma screen at the FedEx Institute of Technology, a building that houses a community of interdisciplinary researchers, at the University of Memphis. As a person approaches the display, MIKI greets them, introduces itself, and offers to be of assistance. The individual can then verbally ask a question related to any of the following topics: (1) events at the FedEx Institute of Technology, (2) research groups housed at the Institute, (3) directions to rooms within the building, (4) people involved in research at the Institute. In answer to the visitor's question, the kiosk responds in a number of different ways. The response might include a verbal answer along with 3-D animations, video presentations, images, or additional audio. Along with the prototype kiosk (Figure 6.1), tools were created that allow for the maintenance and timely update of information presented by the kiosk.

MIKI shares several similarities to the MACK system in that both systems use multiple input channels and that they both provide information on people and groups



FIT 403 - Institute for Intelligent Systems

Room 403 is located on the Fourth Floor. Please use the main elevators directly behind the kiosk. When in the elevator choose 4 as the destination. Upon exiting the elevator turn left. The corridor is the second door on the right.

Fig. 6.1. MIKI: The Memphis Intelligent Kiosk Initiative

and directions to locations at a research site [2, 4]. However, the MACK system relies on rule based grammars alone for speech input. This greatly restricts the scope of questions with which a user can query the system. MIKI is equipped with a standard grammar as well as a statistical based natural language understanding mechanism that reduces the need for rigorous phraseology in the information requests a user presents to the system. We identify this facet of the Intelligent Kiosk as the paramount factor that distinguishes it from other such systems. The MACK system, in contrast, focused primarily on spatial understanding and the way that such an embodied conversational character could make use of the concept of shared space in order to more appropriately give directions to a destination.

MIKI is a collection of different technologies that are all integrated to work seamlessly together. Among these technologies is video processing for face detection, a digital avatar, speaker-independent speech recognition, an advanced graphical user interface (GUI), an array microphone for noise cancellation, a database system, a dynamic question answering system, and a cutting-edge touch panel technology for large displays. We proceed by describing the primary components of the system followed by some of the technical challenges encountered.

6.2 Primary Components

There are several components that comprise the Intelligent Kiosk. A general layout of the major software elements is presented in Figure 6.2.

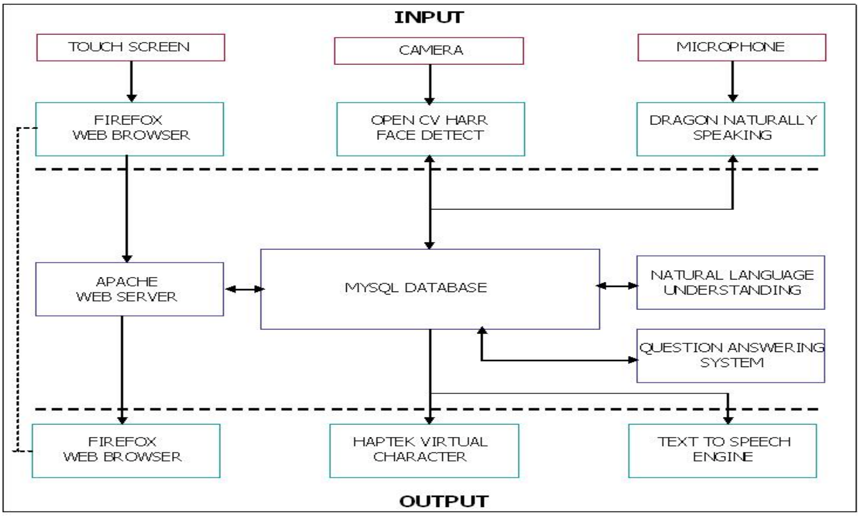


Fig. 6.2. Major software components of the intelligent kiosk

6.2.1 The Hardware

The Intelligent Kiosk resides in the lobby of the FedEx Institute of Technology on a small wall facing the main entrance. The location was chosen for its optimal visibility

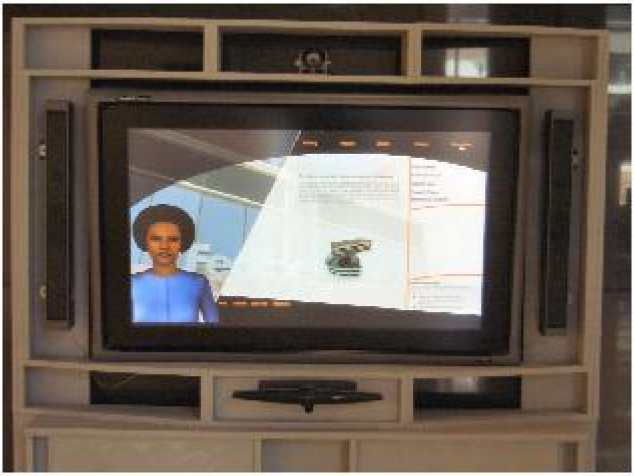


Fig. 6.3. Major hardware components of the intelligent kiosk

to outside visitors. As a visitor enters the building, they see a 50 inch display surrounded by custom-made cabinet. The 50 inch Panasonic® display has been augmented with a touch panel overlay provided by Smart Technologies®. Mounted above the display is a small, FireWire web cam. Just below the display is a somewhat larger Acoustic Magic® array microphone. Both the camera and the array microphone are angled for optimal function for someone standing between 1 and 3 feet in front of the display. Inside the lower part of the cabinet are two Dell® workstations, an Ethernet hub, a KVM switch, and a wireless keyboard and mouse. The two systems are basically identical with 3.2 GHz CPUs, 128 MB AGP video cards, and 2 GB of memory. Both are running windows XP as their operating system. Figure 6.3 depicts the major hardware components of the Intelligent Kiosk.

6.2.2 Speech Recognition

For seamless verbal interaction with visitors seeking information a speaker independent speech recognition system was required. We decided to use the commercially available Dragon NaturallySpeaking developed by Nuance® although it was not designed for speaker independence. This decision was motivated by the fact that we were in the possession of working software to interface with this speech recognition engine through CloudGarden's JSAPI implementation. Therefore, even though we looked into a few other speaker independent speech recognition engines, such as CMU Sphinx, we decided that we would proceed with Dragon NaturallySpeaking version 8 primarily due to time constraints. By carefully restricting the language model for speech recognition we were able to simulate a reasonable quality of speaker independent recognition (more details provided below).

6.2.3 Natural Language Understanding

The natural language understanding module attempts to provide an analysis of the user's utterance in order to determine what action or actions need to be taken. When a system is employed in a limited domain such as the Intelligent Kiosk and has only a limited number of choices of what to say or do next, it need only classify a visitor's utterance rather than completely comprehend every word. MIKI uses two different NLU technologies. These include simple keyword matching and a classification technique based on Latent Semantic Analysis [7, 8].

Classification approaches to NLU include statistical [e.g.,9, 10], information retrieval [e.g.,11, 12] and connectionist approaches [e.g.,13]. Notable among these approaches are those that are based on word co-occurrence patterns such as LSA [7, 8] and HAL [14]. LSA is an attractive approach because it can be trained on untagged texts and is a simple extension to keyword based techniques.

In general, methods for analyzing a large corpus of text, such as LSA, are described as generating "language models" rather than being applied to the specific task of speech recognition. It should be made clear that we are referring to corpus analysis based on large digital texts; this is quite different from the analysis of audio corpuses as is quite common in building dictation grammars for automatic speech recognition systems. What is being proposed here will not alter the audio models used within a speech recognition engine. Instead, a large corpus of text is analyzed in order to create a semantic representation. MIKI attempts to use these semantic representations to categorize incoming utterances as one of the existing grammar rules.

There has been research conducted on the application of this type of corpus analysis to speech recognition, although it is generally assumed that the language models produced would replace or modify those in existing speech recognition engines. Some examples would include Siivola's technique of using a neural network to cluster semantically similar words based on their context [15] and a method created by Gotoh and Renals that automatically generates multiple topic-based language models using a statistical scheme related to singular value decomposition [16]. While the proposed research overlaps with this type of language modeling, it is more akin to information retrieval methods that do text classification. Aside from those already mentioned in the above section, some notable examples would be [17-21].

Latent Semantic Analysis

LSA has been remarkably successful at a number of natural language tasks. In the arena of query-based document retrieval [7, 22], LSA was compared to a large number of research prototypes and commercial systems. LSA was shown to perform as well as the best other method in some trials and as much as 30% better in others, with an average improvement of 16% over the competitors. The next success of LSA was in its modeling of human performance in the Test of English as a Foreign Language (TOEFL) developed by Educational Testing Service [8]. The LSA model answered 64.4% of the questions correctly, which is essentially equivalent to the 64.5% performance for college students from non-English speaking countries. Another of the recent successes is the repeated demonstration that LSA can grade the essays that college students write almost as well as human graders [23-25]. Furthermore, LSA has been very impressive in accounting for (1) the developmental acquisition of vocabulary words, (2) the classification of words into categories, (3) the amount of learning that occurs when students with varying degrees of domain knowledge read a text, and (4) the extent to which sentences in text are coherently related to each other.

To use LSA, one must first develop an LSA space, which acts as a lexicon. The experiments described below were conducted using a combination of Grolier's Encyclopedia and the TASA corpus. The space represents the “meaning” of a word as a vector in a space of K dimensions (where K is typically 100 to 300). The space is built automatically from a training corpus. The corpus consists of a large number of “documents,” where a document could be a sentence, a paragraph or a longer unit of text. From the corpus, one computes a co-occurrence matrix that specifies the number of times that word W_i occurs in document D_j. A standard statistical method, called singular value decomposition, reduces the large WxD co-occurrence matrix to K dimensions. This assigns each word a K-dimensional vector in the space.

Given an LSA space, the similarity (i.e., similarity in meaning, conceptual relatedness, match) between two words is computed as a geometric cosine (or dot product) between the two vectors, with values ranging from -1 to 1. The similarity between two sentences (or longer texts) is computed by first representing the meaning of each sentence as a vector that is the weighted average of the vectors for the words in the sentence, then computing the similarity between those two vectors as a geometric cosine. The match between two language strings can be high even though there are few if any words in common between the two strings. Thus, LSA goes well beyond simple keyword matches because the meaning of a language string is partly determined by the company (other words) that each word keeps.

Connecting LSA and NLU

Understanding spoken speech really requires two very different capabilities. The first is the translation of the sound patterns into textual representations. This process is commonly referred to as speech recognition. This research does not delve into the area of audio to text conversion. We have chosen to use the commercially available Dragon Naturally Speaking™ software package as our speech recognition engine mainly due to the level of developer support available. Like most commercially available speech recognition products, Dragon Naturally Speaking™ can translate audio streams to text using two modes, dictation and grammar, discussed previously.

In a way, our system uses both modes simultaneously. The speech recognition engine is set up to use a list of grammar rules as its primary matching scheme; however, if no rule from this list is matched, the engine will still provide text based on the dictation grammar. For this reason, all the grammar rules used as the basis for semantic classification will still function exactly as originally intended. Our system comes into play only when an existing rule is not matched. In these instances, Latent Semantic Analysis [7, 8] is used to extract and match grammar rules to the text provided by the dictation grammar.

Instantiated grammar rules are represented internally as vectors in LSA space. Here, we need to map an arbitrary utterance to a preexisting rule that then states what knowledge needs to be updated and what actions need to be taken, if any. This is the same task as was performed by the rule-based approach described at the beginning of this section, but, with the LSA approach, the phrases spoken do not necessarily have to match precisely as long as their meanings, expressed as vectors in the LSA space, are similar enough. Unfortunately, there is not a one-to-one correspondence between a grammar rule and a vector in the LSA space. A given utterance may match to

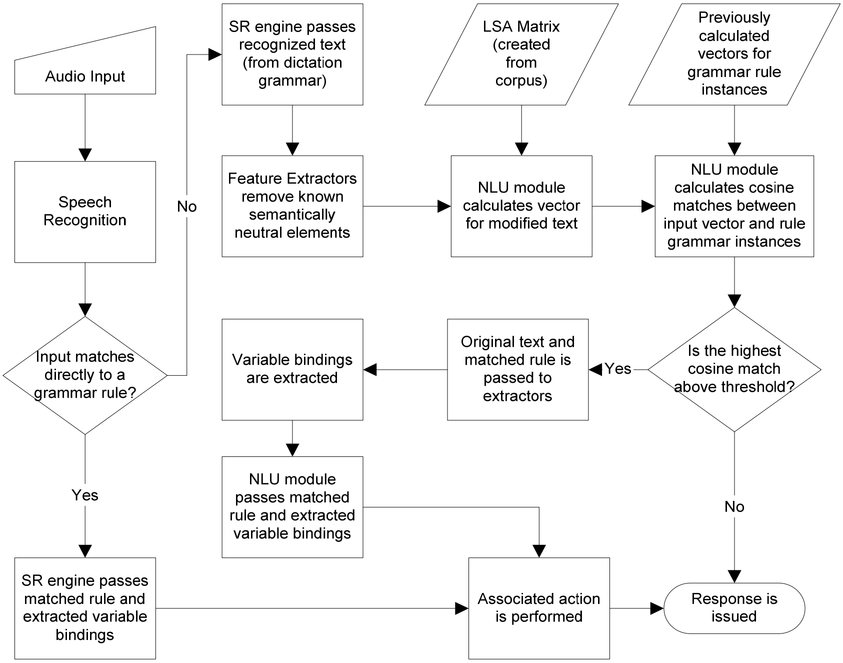

Audio Input
Speech Recognition
Input matches directly to a grammar rule?
No
Yes
SR engine passes recognized text (from dictation grammar)
Feature Extractors remove known semantically neutral elements
LSA Matrix (created from corpus)
NLU module calculates vector for modified text
Previously calculated vectors for grammar rule instances
NLU module calculates cosine matches between input vector and rule grammar instances
Is the highest cosine match above threshold?
Yes
No
Original text and matched rule is passed to extractors
Variable bindings are extracted
NLU module passes matched rule and extracted variable bindings
SR engine passes matched rule and extracted variable bindings
Associated action is performed
Response is issued

Fig. 6.4. Algorithm for natural language understanding

a combination of rules in the grammar. For example, the utterance "when is my next meeting," might match to a pair of rules such as the following:

(1) <whenQuestion> = ‘when is my next <eventType>’
(2) <eventType> = ‘meeting’ | ‘conference’ | ‘luncheon’

Stated loosely, these two rules recognize the words "when is my next" followed by any of the events listed in (2). This very simple combination (not actually used in the research) recognizes three different sentences. LSA would be able to match the utterance against any of the instantiations of the rules, but there is no LSA vector that would match to (1) or (2) individually. Instead, a LSA vector is generated for each possible rule instantiation. Therefore, when an arbitrary utterance is received, its LSA vector is matched against the instantiations of all of the grammar rules currently in context. If a match is found then the system instantiates appropriate data extractors based on the types of variables contained within the matched rule. These extractors are then used to bind values to the necessary variables. These values would be things like a proper name, a date, etc. For the example given above, the <eventType> variable would be bound to “meeting.” Since the grammar rule prescribes which variables need to be bound and their type, regular expressions specific to those types can be used to extract the necessary information from the text. A more complete version of the algorithm is provided in Figure 6.4. The result is that the appropriate grammar rule is “fired” with the correct variables bound just as if the user had uttered the phrase exactly as prescribed in the grammar rule.

6.2.4 People Tracking

The general object detection algorithm consisting of a cascade of boosted classifiers proposed by Viola and Jones [26] was used to detect faces on the basis of Harr-like features. SharperCV, a C# wrapper for the Intel® OpenCV library [27], was used for all video processing. This use of the face detection algorithm provides the ability to not only recognize when someone is approaching the kiosk, but also to count the number of individuals that interacted with the system as opposed to those that walked by. The tracker was based on the assumption that with a large frame rate (30 frames per second) the change in the location of a face from one frame to the next is rather minute. Therefore, a face in the current frame was assigned to the nearest face (based on Euclidian distance) among all the faces in the previous frame as long as the distance was not greater than some predefined maximum threshold.

6.2.5 Graphical Interface

The graphical interface is a combination of two technologies: a standard web-based front end that is the primary interface and a secondary animated character.

Primary Interface

The primary way that individuals can interact with the kiosk is through a web-based front end. It uses a combination of PHP, HTML, and Flash. The PHP and HTML are used to render dynamic pages based on the data stored in the database. The visual transitions between screens as well as direction animations are provided via several Flash scripts. Of particular interest are the animations for directions to various locations in the building. They were generated using the original CAD files of the building. When a user asks for directions to a given room that flash file is loaded. A small ball is illuminated where the user currently is and is then moved along the path that will get that person to their destination. As the floating ball changes floors, that floor becomes visible and the others fade away. In addition to the animation, the directions are also spoken through the avatar and provided in text below the animation.

The Avatar

The animated avatar was created using Haptek's® PeoplePutty software and displayed with the Haptek Player. The programmatic interface through an ActiveX control is fairly straightforward and creates a realistic animated character. The player is free for distribution and has a small footprint both in respect to memory and CPU resources. Using the PeoplePutty software, we were able to create two avatars based on the two students that also provide the voice for their characters. The result is two characters, one male and one female, that look and sound surprisingly like their human models. Despite the fact that the system has a fully functional text-to-speech (TTS) system as shown in Figure 6.2, it was noted that most of the vocal interaction from the agent would be known a priori. As this suggests, all of the vocal output produced by the avatar was prerecorded by the human models. These recordings were then processed using a Haptek tool that tags the recordings with mouth movements. When these recordings are played back through the animated character, the mouth animations of the character match the spoken utterances.

6.2.6 Question Answering

Question answering was accomplished through a combination of mechanisms but primarily involves a naïve use of the database to disambiguate between a number of possible questions being asked. First, each possible type of response is described through an XML file. Because the vocal interaction with the kiosk is intended to be completely open-ended, it is not possible to define each and every question that might be posed to the system. Therefore, each type of response defined in the XML file is defined primarily by what screen services that response requires. For example, all questions regarding general information about an individual will be answered by displaying a detailed personnel page that is dynamically generated based on information from the database. A single element in the XML file describes how to recognize and respond to all inquiries of this type. Each response element then provides the following types of information: (1) the frame name (screen), (2) a key field in the database, (3) a template for an SQL command, (4) text features (as extracted by the speech recognition and/or NLU module) used to match this response to a question, (5) follow-up question types used to disambiguate between several possible answers, (6) a template for a textual response assuming a positive answer was located, and (7) a template for a textual response assuming a positive answer could not be located. With the current system using prerecorded vocal responses, items 6 and 7 above are not used.

The first two items, the frame name and the key field, are used to communicate with the graphical front end. The frame name informs the GUI what screen to display while the ID field contains specific information about which record to display on that screen. For example, a visitor might say, "Can you tell me about John Doe?" The speech recognition/NLU system categorizes this utterance as a request for specific information about John Doe. The question answering system picks this up, matches the question and information provided to a specific response type, and updates a query tracking object with the information provided. It and then binds the variables in the SQL template for that response and runs the database query. Assuming that only one record comes back matching that query, the value in the key field is extracted and a message is sent to the GUI that includes the frame name and the value of the key field for that person's record. The GUI then handles displaying the information on the appropriate screen as well as issuing any messages for actions through the animated avatar.

A distinction should be made between this method and the widely used Artificial Intelligence Mark-up Language (AIML) [28]. Many "chatbots" have been created using AIML as the question-to-answer connection engine. This method essentially maps the text of a question to an appropriate response. What makes this system powerful is its ability to decide between multiple possible answers based on the closeness of the match. MIKI, on the other hand, uses the results of the database search to dictate the final response. As a side note, AIML has been integrated into MIKI as a way of dealing with questions that are not part of the primary database that deals only with information about the FedEx Institute of Technology.

6.3 Technical Challenges

The following section describes the challenges faced during the implementation of the Intelligent Kiosk. Solutions are also presented although some components are still under refinement now that the system is in place.

6.3.1 Component Communication

The various pieces of the kiosk communicate through the database. A table in the database, called the "messages" table, holds all information that is being transmitted between components. The table holds some basic information about each message such as an auto-incremented identifier, the sender, the recipient, the message tag (e.g., LOAD, SPEAK), text data (e.g., a recognized utterance), binary data (e.g., a Hashtable mapping features to values, *first-name=John*), and a timestamp. Any number of components can access the messages table and effectively react to the messages. All they need to know is the tag or tags in which they are interested. This is similar to a shared memory framework.

Using a database as the focus of a communication scheme between components has several advantages. For example, a simple ontology can be used to address senders and information types. This allows the system to be implemented quickly without confusion even among a dispersed group of developers. Adding new resources simply involves a name assignment and the definition of new message tags. Even so, this technique is not significantly different from a standard blackboard model. A central component polls the database on a regular basis and messages are distributed to the appropriate modules. Instead of using specific techniques, such as Galaxy Communicator [29] or other similar systems, it was decided that maximum interoperability would be gained through the use of a method that was common to all of the disparate languages. Even the web-based front end could easily make use of an industry standard database.

The communication scheme is persistent-asynchronous in nature. Message persistence is provided by the messages table in the database. The system is asynchronous because components read and write messages on their own schedules independent of other components. This communication scheme alleviates several problems associated with distributed components. The only thing that each component needs to have a connection to is the database. This is a very simple and well tested procedure that is not specific to a particular language. For instance, the graphical front end is a combination of PHP code running some Macromedia Flash displayed in a Mozilla Firefox web browser. The PHP code has no problems sending or receiving messages from other components to tell it to change screens or display a particular frame. Much of the "back-end" code is written in Java while the virtual agent control software and the tracker is written in C#. Each choice of language for a given component was made based on what core functions were most important and which language supported those functions best. Getting these very different components to talk to each other using some other method like CORBA or SOAP would have been problematic at best and would not have provided any benefit over the database solution. Finally, the database solution already incorporates speed optimizations for the transfer, storage, and logging of data. Logging of the system's internal workings is a simple matter of backing up the messages table.

The use of a database as the medium for sending and delivering messages between components is the central idea that makes the rest of the distributed framework almost trivial. Once the database was in place as the message delivery system, all the components were able to use any method they chose to access content from that database. Each component was then designed as a stand-alone process. In addition to the fact that any component could then be run on any accessible computer, this also facilitated easy testing of components in pseudo-isolation. Testing code manually created messages in the database and then read the resulting messages entered by the component being tested to determine if the test passed or failed.

For the Java and C# components, a generic API to facilitate communication with the database was created to allow for easy implementation of various components. The API contained data structures and algorithms to ensure optimal transparency related to the encoding and decoding of messages to and from the database. Such a system could be said to have scale-up problems if the number of separate components were to become excessively large, but for the foreseeable future of less than one hundred separate modules this system should not have a noticeable degradation in performance.

6.3.2 Speaker Independent Speech Recognition

The commercially available software package Dragon NaturallySpeaking from Nuance® was used for speech recognition. Unfortunately, Dragon NaturallySpeaking was designed for non-speaker independent use. The major motivation for the use of this recognizer over more viable options such as CMU Sphinx was the existence of legacy software using Dragon and a JSAPI implementation from CloudGarden. In order to achieve an acceptable degree of speaker independence we used an untrained speaker profile and replaced the default language model with a much smaller model that consisted of common words that would likely be used in the kiosk domain. To this we added the names of individuals, groups, and some events associated with the FedEx Institute. In other words, the content of the kiosk was used as the majority of the words that the speech recognizer would handle. Additionally, the process of updating the language model was automated by periodically dumping relevant tables (e.g., groups, events, etc) from the database and utilizing the language building tool available as part of the Dragon NaturallySpeaking software suite. Therefore, as people, groups, and events are added, deleted, and modified, the language model is guaranteed to stay consistent.

This method of utilizing a very restrictive language model worked better than was originally expected, but is still unsatisfactory from a performance perspective. Studies are currently under way to determine how well this method performs compared to other methods and recognizers. Ultimately, we are confident that another method or recognition system will be used with better success.

6.3.3 Session Maintenance

One important issue is how and when to start, end, and maintain a session. We would like the kiosk to proactively initiate interaction whenever a person is looking at the screen. The vision component identifies and tracks faces. Based on the size of the face, an approximate distance from the kiosk can be determined. If the face is within about 4 feet of the kiosk, then it is assumed that the person is within the range of interaction. This is used to open a session if one is not currently open, or maintain a currently open session. Voice input and touching of the screen also triggers an opening or maintaining of a session. Finally, sessions are kept open for a few seconds (currently 30 seconds) even if no face is visible and no other input is received. We have found that this works reasonably well. It does not close a session prematurely or keep open a session beyond a reasonable length of time. However, this method does not recognize when a user has disengaged. To appropriately end a conversation, humans usually provide some verbal or body-language cue that he or she is finished. Typically, an “Ok, thanks” might be used or simply, “bye”. MIKI does not currently recognize any of these cues and, therefore, does not satisfactorily terminate the interaction.

6.4 Usability Testing

Increasingly, technology developers have turned to interactive, intelligent kiosks to provide routine communicative functions such as greeting and informing people as they enter public, corporate, retail, or healthcare spaces. A number of studies have found intelligent kiosks—often including avatars—to be usable, with study participants reporting them to be appealing, useful, and even entertaining [30-33]. People are becoming increasingly familiar with avatars populating their informational spaces, virtual and real: Video games and online communities such as Yahoo!®, as well as museums, schools, and other public or institutionalized spaces offer avatars as guides, narrators, and virtual companions [32, 34, 35]. However, as the means for handling information tasks with communication-rich, multimodal interfaces are becoming more feasible, our understanding of the ways in which people select and use the modalities is yet impoverished.

6.4.1 Research Background

Even ordinary tasks that an information kiosk could be expected to handle can pose design and usability issues that complicate standard user-testing efforts. Deeper discussions of information design for multimedia and complex problem-solving have been discussed elsewhere in technical communication research [36, 37], but HCI studies indicate that a successful measure for one aspect of a multimodal system can be dependent on multiple characteristics. For example, social perceptions shaping interaction quality was identified by Stocky and Cassell as contingent on the fit between personality of the avatar and the user, and the consistency of verbal and nonverbal cues across information [38] and suggest increased complexity in capturing valid usability measures.

Research is still at the beginning stages of developing guidelines for successfully incorporating multiple communication modalities, such as written and spoken language content, graphics, an avatar, and speech, into one, seamlessly functioning interface.

Understanding the nature of interactivity is key to designing a useful kiosk. The August spoken dialog system is a kiosk that helps users find their way around Stockholm, Sweden, using an on-screen street map. Much like a sideshow barker at a carnival, August uses speech to elicit a conversation from the user [5, 6], which engages the user with the information at a heightened level than text and map could do. However, this adds another dimension of design considerations: The prevailing belief for speech technology is that users prefer human speech over synthetic speech, or text-to-speech (TTS), but it is impractical to record a complete range of possible responses needed in dynamic voice interfaces with human voice talent. Thus, a combination of human and TTS are typically used. However, Gong and Lai noted in their study that *consistency*, or a seamless match between features of speech quality and perceived personality is key to users' willingness to interact, and ability to comprehend and process information smoothly [39]. Further, Lee and Nass' study shows that increased social presence is perceived when the personality of the speaker and the verbal content match [40].

In terms of making design decisions, what personality type or combinations of characteristics will effectively reach most people in a public setting? One approach to ensuring usability is through user training: The MINNELLI system is designed to interact with bank customers through the use of short animated cartoons that present information about bank services [3]. However, training probably adds an unreasonable burden for a casual user of a general kiosk.

These are but a few of the concerns facing the development team as we set out to design and build an intelligent kiosk, but it was clear that as the "reach" of multimodal interfaces increasingly extends beyond the immediate screen of a monitor, effective design must include consideration of the physical presence and fit of the object, in this case, a kiosk, and the environment in which it resides, and is used.

6.4.2 Design Questions

In designing the kiosk, a number of questions emerged that drove our design choices:

When faced with a rich, multi-modal interface, how do people select and use modalities for basic information seeking tasks?

How do people select and use modalities for location-seeking tasks? Can richer graphical information (3D) help people navigate the building more successfully than limited graphical information (2D)?

What makes an appealing avatar? Do different people respond differently to being greeted and given information by avatars of different gender, ethnicity, and perceived social standing?

Along with the richness of offering an avatar, come many design decisions about both the obviously visible as well as the subtle features of the humanlike companion: physical appearance including gender; skin color; hair texture, length, color, and style; eye shape and color; body type, and even how much of the body to include. In addition, subtle features of human behavior that informs our "reading" of the verbal messages we convey when in person, such as eye gaze, body movements, voice quality, accent, and pacing.

When given an avatar "greeter," complex information delivery and use is a technical communication issue addressed as an increasingly important issue as information is displayed, shaped, and delivered via multiple modalities. A usability evaluation tested 38 users' abilities to accomplish the most common tasks MIKI was designed to support. Testing methodology, results, and discussion follow.

6.4.3 Methodology

The usability test employed a mixed measures design (see Table 6.1).

Table 6.1. Design Factors

Factor	Level Names	# Participants
Gender	Female	23
	Male	15
Avatar Persona	Khadejah (K)	21
	Vince (V)	17
Discipline	Humanities	28
	Engineering	10
Task Order	Place→Event→Person	11
	Place→Person→Event	0
	Event→Place→Person	5
	Event→Person→Place	4
	Person→Place→Event	10
	Person→Event→Place	8

Tasks & Measures

Each participant was verbally instructed to complete three tasks with the Intelligent Kiosk:

1. Find a person
2. Find a place
3. Find an event

These tasks were representative of the most commonly requested information by visitors to the FedEx Institute of Technology.

Seven measures were used in usability testing, three scaled instruments: (1) After Task Questionnaire (ATQ), (2) Usability Scale for Speech Interfaces (USSI), (3) Post-Scenario System Usability Questionnaire (PSSUQ); three observational measures: (4) Task Completion Measures, (5) Observed Usability Problems, (6) Observed Interaction Preferences; and (7) Qualitative Interviews, using a cued recall technique.

Participants

Thirty-eight participants (approximately 60% female and 40% male) were recruited from two summer communication course sections at the University of Memphis (see Table 6.2). Ten participants were drawn from a course comprised of engineering students, a technology-intensive discipline, and twenty-eight were drawn from a section of humanities majors, a relatively technology-non-intensive discipline. Participants received extra course credit for their participation in the study.

Table 6.2. Participant Makeup by Gender

Discipline	Male	Female	Total
Engineering	8	2	10
Humanities	7	21	28

Usability Test Measures

Measurement consisted of a variety of observational measures and rating scales, as well as participant responses to interview items. The measures presented to each participant were:

1. After Task Questionnaire (ATQ) – The ATQ is a validated 3-item, 7-point scale that measures the user's immediate satisfaction of the task just completed [41] Users filled out one for each of the three tasks. To complete the scale, participants rated their relative agreement with each item:

I am satisfied with the ease of completing this task

I am satisfied with the amount of time it took to complete this task

I am satisfied with the support information (online help, messages, documentation) when completing this task

2. The Usability Scale for Speech Interfaces – This scale is a 25-item, 7-point measure that assesses the usability of speech technologies [42]. It uses four factor scores—U**ser Goal Orientation, Speech Characteristics, Customer Service Behavior, and Verbosity—**with 6 – 8 items for each factor:

Sample items
The Intelligent Kiosk made me feel like I was in control.
I could find what I needed without any difficulty.
The avatar's voice was pleasant
This avatar used everyday words.
The avatar spoke at a pace that was easy to follow
The messages were repetitive.
The avatar was too talkative.

Participants filled out the scale after all three tasks were completed, rating their relative agreement with each item.

3. Post-Scenario System Usability Questionnaire (PSSUQ) – The PSSUQ is a validated 16-item, 7-point scale that measures usability [43]. It provides three factor scores—**System Usefulness**, **Interface Quality, and Information Quality**—as well as an overall usability score. Participants filled out the scale after all three tasks were completed, rating their relative agreement with each item.

Sample items
I am satisfied with how easy it is to use the Intelligent Kiosk.
It was simple to use this system.

4. Task Completion – For each task, the evaluators recorded time on task, and whether or not the participant successfully completed the task.

5. Observed Usability Problems – As each participant completed each task, two evaluators observed participant behavior to describe and record any usability problems encountered during the task. After the evaluation, each usability problem was ranked according to severity:

1 = no usability problems observed

2 = mild confusion <1min independent task completion

3 = confusion >1min independent task completion

4 = confusion with task stoppage recovery using provided supports (e.g., help)

5 = task failure or abandonment

6. Interaction Preference – During each task, evaluators recorded whether or not participants used either the graphic user interface (GUI), the speech user interface (SUI), or both.

7. Qualitative Interview Items – Two evaluators interviewed participants, prompting them by showing them problem screens noted in 5. Observed Usability Problems. Cued recall interviews helped flesh out details of participants' likes/dislikes, their interaction preferences, places they thought necessary information was missing, avatar-related, and other design changes. When appropriate, they were asked to describe the perceived source of any confusion or usability problems they encountered during task completion.

6.4.4 Usability Test Results

An ANOVA with order cast as a between subjects variable indicated a non-significant effect (p > 0.05); therefore, order effects were not present and were not considered in subsequent analyses.

After Task Questionnaire (ATQ)
Participants rated their satisfaction levels for the Find a Person and Find an Event tasks relatively positively with the Find a Place task rated below the midpoint of the scale.

A three-way mixed model repeated measure ANOVA was performed with participant ratings on the ATQ for each task as the within-subjects factor and the participant's gender, discipline (engineering or humanities), and the avatar persona (Khadejah or Vince) as between-subject factors. The main effect of task was statistically significant, $F(2, 62) = 43.077$, $MSe = 1.523$, $p < .01$. Bonferroni post hoc tests revealed that

Table 6.3. After Task Questionnaire (7-pt scale)

Task	Mean	SD
Find Person	6.65	0.50
Find Place	3.73	1.16
Find Event	6.90	0.32

participant ratings after locating a person and an event were on par and significantly higher ($p < .01$) than the ratings associated with finding a place. The main effects for the between subject factors including participant gender, participant discipline, and avatar persona were not statistically significant ($p > 0.4$). Additionally, higher order interactions between the ATQ and the other between-subject factors were not statistically significant.

The Usability Scale for Speech Interfaces

Table 6.4 provides the mean and standard deviation for each of the four factors of the Usability Scale for Speech Interfaces.

Table 6.4. Usability Scale for Speech Interface

Speech Categories	Mean	SD
User Goal Orientation	5.28	1.46
Speech Characteristics	5.04	1.74
Customer Service Behavior	6.02	1.25
Verbosity	3.52	1.91

A three-way mixed model repeated measure ANOVA was performed with participant ratings on the four factors of the USSI as the within-subjects factor and the participant's gender, discipline (engineering or humanities), and the avatar persona (K or V) as between-subject factors. The main effect of USSI factor was statistically significant, $F(3, 93) = 13.696$, $MSe = 1.752$, $p < .01$ (partial $\eta^2 = .306$), suggesting that participant ratings significantly differed across the four factors of the USSI. Main effects for the between subject factors including participant gender, participant discipline, and avatar persona were not statistically significant ($p > 0.05$).

Higher order interactions between the USSI and participant gender and discipline were not statistically significant ($p > .05$). However, a statistically significant interaction between participant ratings on the USSI and gender was discovered, $F(3, 93) = 7.692$, $p < 0.01$ (partial $\eta^2 = .199$). Participants of both genders provided similar ratings with respect to the User Goal Orientation factor; however, ratings by males for the Speech Characteristics and Customer Service Behavior factors were quantitatively lower than their female counterparts (see Figure 6.5). Additionally, females provided lower Verbosity ratings than males.

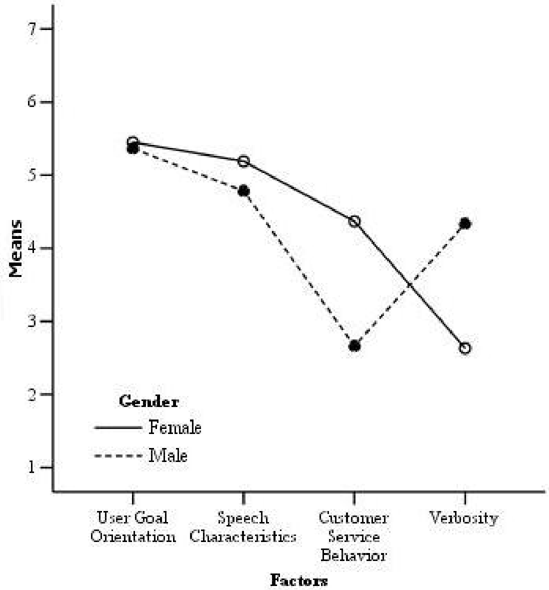


Fig. 6.5. Means for participant ratings on the USSI segregated by participant gender

Post-Scenario System Usability Questionnaire (PSSUQ)

A three-way mixed method repeated measure ANOVA was performed, with participant ratings on the PSSUQ as the within-subjects factor, and the participant's gender, discipline (engineering or humanities), and the avatar persona (K or V) as between-subject factors. Participant ratings across the 3 factors of the scales were similar (see Table 6.5).

Table 6.5. Post-Scenario System Usability

Interface Categories	Mean	SD
System Usefulness	5.54	1.48
Interface Quality	5.56	1.59
Information Quality	5.04	1.79
Overall	5.36	1.64

Main effects for the three between subject factors including participant gender, participant discipline, and avatar persona were not statistically significant (p > 0.05). Additionally, higher order interactions between participant ratings on the PSSUQ and the other two between subject factors were not statistically significant (p > 0.05).

Task Completion

Overall, task completion rates were high (see Figure 6): All participants successfully completed the Find a Person task, and only one person failed to find an event. The lowest completion rate was shown by the Find a Place task, in which 5 of 38

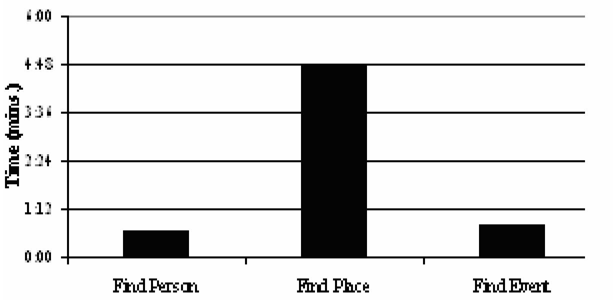


Fig. 6.6. Task Completion Times

participants failed. Additionally, completion time for the Find a Place task was also elevated in comparison to the other two tasks.

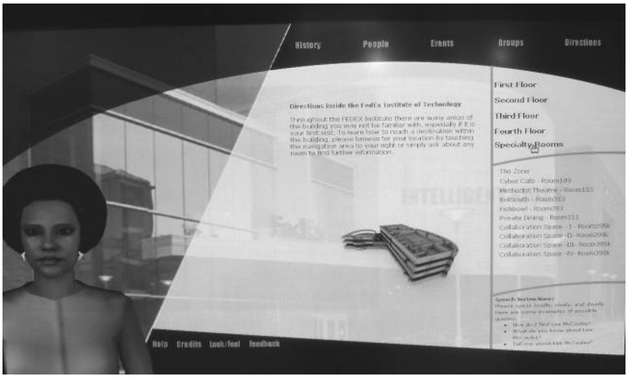
Observed Usability Problems

By far, people exhibited the most problems trying to accomplish the Find a Place task. We observed a number of problems contributed to this result:

First, tab labels were unclear to most people. Further unclear language for lists, such as Specialty Rooms misled 33 out of 38 participants.

Once the language failed them in the first two to three steps, twenty-nine of 38 participants were reduced to cycling through links to the four lists for the building's four floors. However, the lists did not contain enough information, giving room numbers, but not names of centers, further frustrating participants.

Navigational expectations were unmet for many: Fifteen of the 38 participants looked for a direct link from the Directions screen to the center's home page. The text and navigational problems were then compounded when they failed to find adequate help (17/38), and, finally, when 22 participants turned to the SUI, the voice recognition failed 14 of them.



Directions inside the FedEx Institute of Technology

Throughout the FEDEX Institute there are many areas of the building you may not be familiar with, especially if it is your first visit. To learn how to reach a destination within the building, please browse for your location by touching the navigation area to your right or simply ask about any room to find further information.

The Zone
Cyber Cafe - Room109
Methodist Theatre - Room103
BellSouth - Room313
Fishbowl - Room203
Private Dining - Room111
Collaboration Space - I - Room299a
Collaboration Space -II- Room299b
Collaboration Space -III- Room399a
Collaboration Space -IV- Room399b

Speech Instructions:
Please speak loudly, slowly, and clearly. Here are some examples of possible queries:

- How do I find Lee McCauley?
- What do you know about Lee McCauley?
- Tell me about Lee McCauley?

Fig. 6.7. MIKI Directions screen showing 3D floor plan of the building. Floor plan continually rotates as long as screen is visible.

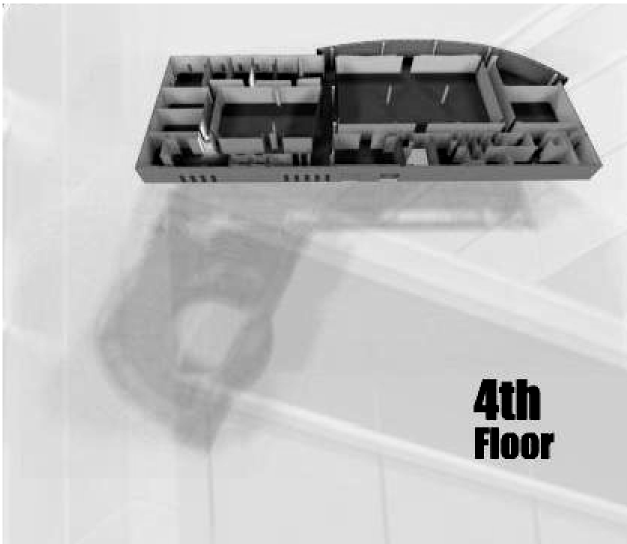


Fig. 6.8. Close-up of 3D floor plan

Finally, observational notes indicated that participants did not find the 3D animated floor plan helpful (see Figures 6.7 and 6.8). One participant's comment summarizes well the observed participant experience: "Too many things going on with directions—animation, voice, text—confusing."

Interaction Preference

A count of participants who used the GUI, SUI, or both interfaces showed a strong preference for the GUI interface, as shown in Table 6.6. Notably, for the Find a Place task, the majority of participants used both interfaces, possibly due to the number of usability problems encountered during this task.

When asked which interface they prefer, participants also expressed a strong preference for the touch screen/GUI (84% of participants), as compared with the voice interface (5%) and both interfaces (11%).

Table 6.6. Interaction Preference

Task	GUI	SUI	Both
Find Person	33	0	5
Find Place	1	0	37
Find Event	33	4	1

6.5 Future Directions

The current version of the intelligent kiosk serves as a useful prototype for usability analyses and as a test bed for general issues involving speech recognition in noisy environments, natural language understanding from speech, and question answering from incomplete queries. The software which is an amalgamation of Java, C#, C++, HTML, and PHP is modular and extendable with minimal component interaction thus yielding very low coupling and high cohesion. Since each component executes as a separate resource (or process) components can be migrated to additional hardware on the fly. This opens up an interesting research forum for issues related to dynamic resource allocation and recovery from partial failure.

The Intelligent Kiosk is still under development from a number of different directions. The internal workings are being updated to increase the speech recognition accuracy and cosmetic changes are being made to the interface in order to increase user interaction and satisfaction with their experience. From the perspective of speech interaction, a different speech recognition engine will be installed that is specifically designed for speaker-independent recognition.

From the usability tests, MIKI exemplifies much of good interface design, primarily the quality of the graphics, screen layout, and organization of most of the information. However, a number of lessons were learned in this project that can be applied to further development of MIKI and of other kiosk interfaces that are intended to address information tasks of varying complexity for a public audience.

The data indicated that participants needed more time and encountered more problems with the Find a Place task. This finding appeared to be at least partly due to the lack of visual orientation of the 3D building floor plan. Adding clearly marked start and end points would help anchor the image of the building to the visitor's sense of the physical context—both the immediate physical surroundings of the lobby and kiosk area, as well as the building beyond.

In addition, although the spinning 3D floor plan may have added visual interest to the screen, the user could not control the animation—either the speed of spinning, or the capability to stop, reverse, or zoom in. Thus, the user could not view at an individual pace, or stop the spinning to study details. We believe that any potential to enhance the user's sense of the building by showing it from all angles was at best, diminished. We do not know how much the added visual busy-ness may have impaired the user's ability to understand the information or, ultimately, to accomplish the task. Would a 2D image been easier to grasp than 3D? What kind of cognitive load did this bear on the user as he or she was juggling other processing tasks? Accordingly, next phase development will include adding points of reference (e.g., "You are here" and "Your destination is here"), further exploration of 2D vs. 3D, and value added by the ability to spin the image.

Finally, we do not know how much better participants would have performed this task had they had some type of aide memoire of the information once they found the correct directions. We did not specifically measure how long they could retain the directions in memory, but our informal observations of the participants' confusion after leaving the kiosk suggest this may have been the case. Possibly a printout or a downloadable set of directions would be necessary to fully help users with this task.

Making the speech feature more robust is clearly necessary, and a more focused analysis of the advantage of speech for certain information tasks would be a first step. A more fine-grained application of speech technology in which we target the points at which people turn to speech of the other modalities offered, would be most helpful would not only alleviate the burden on a whole-system speech integration, but would take better advantage of the strengths of offering information through speech, as opposed to pictures or text.

As far as developing best practices for using avatars in multimodal interfaces, much more work remains on the interplay of persona features. Initial studies into the cultural impact of our choices of avatar gender, ethnicity, social standing, and culturally shaped behaviors such as eye gaze, not only mediate the interaction and perceived quality of the communication [44], but also convey paralinguistic information shaping the message and perpetuating cultural attitudes [45].

With additional focused research and development, a robust, interactive information kiosk can be successfully deployed in a number of different domains. Some examples include retail outlets such as malls and "big-box" type stores. Additionally, healthcare institutions and corporate office buildings can also make use of this type of kiosk. The key limitation, at this point, rests in the general applicability of speech recognition for a broad audience and a diverse conversational domain. Our research also points to the difficulty of using speech as an interface medium within a public space. Further research needs to be conducted to better quantify this effect.

Finally, additional installations are being pursued. These might include retail stores, local corporate office buildings, or healthcare institutions. Based on current experience, we would expect these installations could be implemented within three months primarily due to artistic customization.

Acknowledgements

This project was funded by FedEx Institute of Technology Innovation Grants. In addition, many people have contributed to the creation of the Intelligent Kiosk. We would like to acknowledge James Markham from the Intelligent Environments Research Lab (IE-Lab) and Lucas Charles, Austin Hunter, Kristin Mudd, Susan Popham, and Michael Schmidt from the Center for Multimedia Arts (CMA) at the University of Memphis. We would also like to acknowledge Scansoft for donating the Dragon NaturallySpeaking v8.0 SDK and Smart Technologies Inc. for providing us with an Actalyst touch screen panel at no cost.

References

[1] Stephanidis, C., Salvendy, G., Akoumianakis, D., Bevan, N., Brewer, J., Emiliani, P.L., Galetsas, A., Haataja, S., Iakovidis, I., Jacko, J., Jenkins, P., Karshmer, A., Korn, P., Marcus, A., Murphy, H., Stary, C., Vanderheiden, G., Weber, G., Ziegler, J.: Toward an Information Society for All: An International R&D Agenda. International Journal of Human-Computer Interaction 10, 107–134 (1998)

[2] Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, H.: MACK: Media lab Autonomous Conversational Kiosk. In: Imagina 2002, Monte Carlo (2002)

[3] Steiger, P., Suter, B.A.: MINELLI - Experiences with an Interactive Information Kiosk for Casual Users. In: UBILAB 1994, Zurich (1994)

[4] Stocky, T., Cassell, J.: Reality: Spatial Intelligence in Intuitive User Interfaces. In: Intelligent User Interfaces, San Francisco, CA (2002)

[5] Gustafson, J., Lindberg, N., Lundeberg, M.: The August spoken dialogue system. In: Eurospeech 1999 (1999)

[6] Gustafson, J., Lundeberg, M., Liljencrants, J.: Experiences from the development of August - a multimodal spoken dialogue system. In: IDS (1999)

[7] Dumais, S.T.: Latent semantic indexing (LSI) and TREC-2. In: Harman, D. (ed.) National Institute of Standards and Technology Text Retrieval Conference (1994)

[8] Landaur, T.K., Dumais, S.T.: A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240 (1997)

[9] Charniak, E.: Statistical Language Analysis. Cambridge University Press, Cambridge (1993)

[10] Sanker, A., Gorin, A.: Adaptive language acquisition in a multi-sensory device. IEEE Transactions on Systems, Man and Cybernetics (1993)

[11] Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., Group, T.R.: Using Latent Semantic Analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments 8, 149–169 (2000)

[12] Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A.C.: Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In: Proceedings of Artificial Intelligence in Education 1999, pp. 535–542. IOS Press, Amsterdam (1999)

[13] Miikkulainen, R.: Subsymbolic case-role analysis of sentences with embedded clauses. Cognitive Science 20, 47–74 (1996)

[14] Burgess, C., Livesay, K., Lund, K.: Explorations in Context Space: Words, Sentences, Discourse. Discourse Processes 25, 211–257 (1998)

[15] Siivola, V.: Language modeling based on neural clustering of words. In: IDIAP-Com 2002, Martigny, Switzerland (2000)

[16] Gotoh, Y., Renals, S.: Topic-based mixture language modeling. In: Natural Language Engineering, vol. 6 (2000)

[17] Baker, L.D., McCallum, A.K.: Distributional clustering of words for text classification. In: SIGIR 1998, Melbourne, Australia (1998)

[18] Bellegarda, J.R., Butzberger, J.W., Chow, Y.L., Coccaro, N., Naik, D.: Automatic Discovery of Word Classes Through Latent Semantic Analysis. In: EUSIPCO 1996 Signal Processing VIII, Theories and Applications (1996)

[19] Bellegarda, J.R., Butzberger, J.W., Chow, Y.L., Coccaro, N., Naik, D.: A Novel Word Clustering Algorithm Based on Latent Semantic Analysis. In: ICASSP 1996 (1996)

[20] Khudanpur, S., Wu, J.: A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition. In: ICASSP 1999, Phoenix, AZ (1999)

[21] Kuhn, R., De Mori, R.: A cashe-based natural language model for speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 570–583 (1990)

[22] Deerwester, S., Dumais, S.T., Fumas, G.W., Landaur, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990)

[23] Foltz, P.W., Britt, M.A., Perfetti, C.A.: Reasoning from multiple texts: An automatic analysis of readers' situation models. In: Proceedings of the 18th Annual Conference of the Cognitive Science Society, pp. 110–115, Erlbaum, Mahwah, NJ (1996)

[24] Landaur, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)

[25] Wolfe, M.B.W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K.: Learning from text: Matching readers and texts by latent semantic analysis. Discourse Processes 25, 309–336 (1998)

[26] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. CVPR (2001)

[27] The Intel Open Source Computer Vision Library. Intel Corp., vol. 2006 (2006)

[28] Wallace, R.J.: The Elements of AIML Style. In: ALICE A. I. Foundation (2003)

[29] Galaxy Communicator, The MITRE Corporation (2006)

[30] Cavalluzi, A., Carolis, B.D., Pizzutilo, S., Cozzologno, G.: Interacting with embodied agents in public environments. In: AVI 2004 (2004)

[31] Christian, A.D., Avery, B.L.: Speak out and annoy someone: Experiences with intelligent kiosks. In: Proceedings of Computer Human Interaction (CHI) 1998 (2002)

[32] Guinn, C., Hubal, R.: An evaluation of virtual human technology in informational kiosks. In: 6th International Conference on Multimodal Interfaces, State College, Pennsylvania (2004)

[33] Mäkinen, E., Patomäki, S., Raisamo, R.: Experiences on a multimodal information kiosk with an interactive agent. In: NordiCHI, Ärhus, Denmark (2002)

[34] Yahoo!® Canada Avatars, vol. 2006 (2006)

[35] Roussou, M., Trahanias, P., Giannoulis, G., Kamarinos, G., Argyros, A., Tsakiris, D., Georgiadis, P., Burgard, W., Haehnel, D., Cremers, A.B., Schulz, D., Moors, M., Spirtounias, E., Marianthi, M., Savvaides, V., Reitelman, A., Konstantios, D., Katselaki, A.: Experiences from the use of a robotic avatar in a museum setting. In: Proceedings of the 2001 Conference on Virtual Reality, Archeology, and Cultural Heritage, Glyfada, Greece (2001)

[36] Albers, M.: Design considerations for complex problem-solving. In: STC Proceedings (2002)

[37] Hackos, J.T., Hammar, M., Elser, A.: Customer partnering: Data gathering for complex online documentation. Com Tech Services (1997)

[38] Stocky, T., Cassell, J.: Shared reality: Spatial intelligence in intuitive user interfaces. In: Intelligent User Interfaces, San Francisco, CA (2002)

[39] Gong, L., Lai, J.: Shall we mix synthetic speech and human speech? Impact on user's performance, perception, and attitude. In: Proceedings of SIGCHI, vol. 3, pp. 158–165 (2001)

[40] Lee, K.M., Nass, C.: Desinging social presence of social actors in human computer interaction. In: Proceedings of Computer Human Interaction, vol. 5, pp. 289–296 (2003)

[41] Lewis, J.R.: IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction 7, 57–78 (1995)

[42] Polkosky, M.: Toward a psychology of speech technology: Affective responses to speech-based e-service. Doctoral dissertation: University of South Florida (2005)

[43] Lewis, J.R.: Psychometric evaluation of the PSSUQ using data from five years of usability studies. International Journal of Human-Computer Interaction 14, 463–488 (2002)

[44] Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., Sasse, M.A.: The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. Computer Human Interaction (2003)

[45] Zdenek, S.: Just roll your mouse over me: Designing virtual women for customer service on the web. Technical Communication Quarterly 16 (2007)

Author Index

Editors

Professor Dr. Nadia Magnenat-Thalmann has pioneered research into Virtual Humans over the last 25 years. She obtained several Bachelor's and Master's degrees in various disciplines (Psychology, Biology, Computer Science and Chemistry) and a PhD in Quantum Physics from the University of Geneva in 1977. She moved to the University of Geneva in 1989 where she recreated MIRALab. She has received several scientific and artistic awards for the research and the films she has directed. More recently, she has been elected to the SWISS ACADEMY OF TECHNICAL SCIENCES, selected as a pioneer of Information Technology at the HEINZ NIXDORF MUSEUM'S Electronic Wall of Fame in Germany (www.hnf.de). She has published more than 400 papers in the area of Virtual Humans, Mixed and augmented reality.

She is editor-in-chief of the VISUAL COMPUTER JOURNAL published by Springer Verlag, and co-editor-in-chief of the COMPUTER ANIMATION AND VIRTUAL WORLDS journal published by Wiley. She has also served on a number of senior academic executives' positions, including Vice-Rector of the University of Geneva Switzerland. (www.miralab.unige.ch;thalmann@miralab.unige.ch)

Dr. Nikhil Shripal Ichalkaranje is Senior Technology Advisor in the Australian Department of Broadband, Communications and the Digital Economy. He is also an adjunct Senior Research Fellow at the University of South Australia. His research interests include artificial intelligence, computer communications/networking and robotics. Nikhil holds a Masters Degree in Computer Network Systems from Swinburne University of Technology Melbourne and a PhD in Computer Systems Engineering from the University of South Australia Adelaide. Nikhil has co-edited 4 books along with several publications in the area of artificial intelligence and their applications.

Professor Dr. Lakhmi C. Jain is a Director/Founder of the Knowledge-Based Intelligent Engineering Systems (KES) Centre, located in the University of South Australia. He is a fellow of the Institution of Engineers Australia.

His interests focus on the artificial intelligence paradigms and their applications in complex systems, art-science fusion, virtual systems, e-education, e-healthcare, unmanned air vehicles and intelligent agents.